arXiv:2303.05911v2 [cs.LG] 4 Jun 2023

终身机器学习势

MarcoEckhoff* 和 Markus Reiher†苏黎世联邦理工学院，化学与应用生物科学系，*Vladimir-Prelog-Weg 2*，*8093* 苏黎世，瑞士。（日期：2023年5月11日）

基于高精度量子化学数据训练的机器学习势（MLPs）能够保持高准确率，同时只需很少的计算资源。但缺点是，每个单独的体系都需要专门训练。近年来，大量机器学习势从头开始训练，因为在通常情况下，学习额外数据需要重新使用所有数据进行训练，以免遗忘先前获取的知识。此外，机器学习势最常用的结构描述符无法高效表示大量不同的化学元素。在这项工作中，我们通过引入元素包容型原子中心对称函数（eeACSFs）来解决这些问题，这种函数将结构性质与元素周期表中的元素信息结合起来。这些 eeACSFs 是我们开发终身机器学习势（lMLP）的关键。通过不确定性量化，可以突破固定的预训练机器学习势的限制，构建能够持续适应的 lMLP，同时确保达到预先设定的准确率。为了将 lMLP 的适用范围扩展到新体系，我们采用持续学习策略，使其能够针对连续数据流中的新数据自主进行在线训练。对于深度神经网络的训练，我们提出了持续弹性（CoRe）优化器，以及依赖于数据回放、参数正则化和模型架构的增量学习策略。

Key词语：终身机器学习、持续弹性（CoRe）优化器、元素包容型原子中心对称函数 - Cen高维神经网络势、不确定性量化

1. 引言

要预测并理解原子系统的性质和反应性，了解其势能面不可或缺。势能面由电子能随原子核位置变化的函数给出，可通过密度泛函理论（DFT）等电子结构方法或近似波函数理论 [1, 2]获得。这些方法通常具有普适性，对许多体系都能给出准确结果，但即使采用最先进的方法，在扩展模拟中仍会产生很高的计算需求 [3–5]。经验力场不再依据玻恩–奥本海默近似，针对每一种原子核构象显式计算电子能，而是采用近似的（简单的）解析表达式来描述势能面。因此，它们无需显式求解量子力学方程，从而能够高效地进行原子模拟。然而，力场表达式的准确性有限，通常也无法普遍适用于所有化学体系[6–9]。

相比之下，机器学习势（MLPs） [10–15] 能够保留电子结构方法的高准确率，同时计算成本却与力场相当。机器学习势并非基于物理近似，而是依靠高度灵活的数学表达式来获得解析形式的势能面。这些表达式的参数通过训练获得，训练所用的数据是电子结构参考数据，包括化学结构及其相应的能量和原子力。第一代机器学习势仅适用于低维系统 [16], ；第二代机器学习势以高维神经网络势（HDNNP）的形式问世 [17–19] ，使得人们能够在纳秒时间尺度上对含有数万个原子的系统进行原子模拟，其准确率与第一性原理方法相近。近年来，人们还提出了多种其他类型的机器学习势，例如神经网络势 [20–23], 、高斯近似势 [24], 、矩张量势 [25], 等。第二代机器学习势依赖于大部分原子相互作用的局域性 [19]。然而，第三代机器学习势加入了对长程相互作用的描述 [24, 26], 第四代机器学习势甚至能够描述超出原子相互作用截断球范围的非局域相互作用 [27, 28] 。

原子结构参考数据需要充分表征后续模拟将探索的构象空间。原因在于，机器学习势旨在对已学习的原子相互作用进行插值，但其在远超训练所覆盖的原子环境空间时，外推能力有限 [12, 18]。例如，在水上训练的多层感知机用于气态氢分子和氧分子时将无法奏效。对于大多数系统而言，由于原子系统通常可访问的构象空间极其庞大，系统性构建方法并不可行；尤其是在考虑各种大振幅运动或化学反应时更是如此。为获得合理的参考构象，人们设计了迭代式主动学习协议 [29–33]。在此类协议中，先在原子模拟中验证初步的多层感知机，以识别重要但尚未

* marco.eckhoff@phys.chem.ethz.ch
† mreiher@ethz.ch

探索到的未知构象。随后使用参考方法重新计算这些构象，并将其加入参考数据中，以训练出改进后的多层感知机。然而，即使针对特定应用，也很难确保参考数据集完整，因为某些构象可能极少出现。此外，对于同一系统，不同的科学目标和用途可能会关注其构象空间中的不同区域，因此需要不同的参考构象。

当前的机器学习势都是孤立地进行学习，也就是说，模型在预先定义的数据上训练，随后应用于模拟，而不会利用已学到的知识为未来的机器学习势提供帮助。因此，尽管具有通用的函数形式，机器学习势仍只能可靠地应用于特定的构象空间——该空间可能涵盖多种化学式和构型——而且其在已学习的原子环境空间之外的迁移能力十分有限 [8, 9]。因此，人们构建了大量单用途机器学习势 [34–42]。近期研究尝试通过利用规模庞大且涵盖广泛的训练数据集[43, 44], 来克服这一局限，但未来的应用仍可能需要不同的参考数据。

要实现通用解决方案，机器学习势需要能够基于已获取的知识进行适应性学习，以掌握新的化学体系。显然，可以简单地扩展当今机器学习势的训练数据集，在主动学习过程中纳入其他体系。然而，新数据集或扩展后的数据集的训练过程通常会从头开始，并随机初始化机器学习势参数。即便使用先前获得的参数，训练通常也不会区分新旧数据点。因此，每次添加新数据时，都需要将所有数据点放在静态批次中同时训练。由此，随着旧数据重复训练量的增加，学习更大的数据集会变得越来越低效。这一趋势与人脑的运作方式背道而驰；人脑能够基于已掌握的专长，高效地学习更多知识 [45–47]。

这种增量学习能力在机器学习领域被称为终身学习或持续学习 [46, 47] ，机器学习势也需要具备这种能力。终身机器学习意味着模型持续获取并微调知识。对于机器学习势而言，这属于单增量任务场景，因为新的化学结构会依次加入，但必须区分所有已遇到结构的能量 [48]。终身学习是机器学习面临的一项挑战，因为从连续信息流中进行增量学习通常会导致灾难性遗忘和干扰 [49, 50]。机器学习势的参数会根据新数据进行调整，而对旧数据的记忆则逐渐消退。

原则上，利用已有知识甚至可以减少所需的训练数据量，并更高效地学习更复杂的任务。根据互补学习系统理论， [47], 学习基于情景记忆和泛化。情景记忆能够快速学习任意信息，而泛化则是一个缓慢的学习过程，用于构建知识结构。因此，情景记忆类似于训练数据集，而多层感知机参数则经过训练以实现数据泛化。因此，参数需要具有可塑性，以整合新知识；但同时也需要保持稳定，以避免遗忘 [45–47]。近年来，已有多种算法被开发出来，以实现这种稳定性-可塑性平衡并缓解灾难性遗忘；例如 ExStream [51], 梯度情景记忆 [52], 突触智能 [53], AR1 [48], 渐进式神经网络[54], 和增长式双记忆 [55]。这些方法依赖于对选定训练数据的回放、对机器学习参数进行正则化，和/或机器学习模型的架构。

对于通用型或普适型机器学习势，情况会更加复杂，因为随着化学元素数量增加，其计算成本的不利扩展特性会进一步造成阻碍。这一特性困扰着大多数常用的机器学习势描述符 [9, 56, 57] ，例如原子中心对称函数（ACSFs） [58], 、原子位置的平滑重叠 [59], 以及许多其他方法 [60, 61]——图表示 [62–64]除外。其原因在于结构描述符的数量迅速增加，因为这些描述符是针对每一种元素组合独立构建的。一些研究尝试通过在结构描述符中纳入一个额外的、依赖于元素的权重 [65–67]，将元素信息与结构信息结合起来，以克服这一问题。尽管已有一些成功应用，例如在加权原子中心对称函数 [67] 中采用原子序数，但这种方法仅适用于原子序数相近的元素。否则，轻原子（如氢）的贡献会被重原子（如碘）的贡献掩盖（例如，在角度贡献中，系数为 $53 \cdot 53$ ）。

在这项工作中，我们针对不同的原子中心对称函数，利用基于不同性质的额外权重项，从而减少描述符对特定元素的偏向，并提高其对不同构象的判别能力。化学直觉基于分子结构中的规律以及元素周期表中的趋势。为了利用这些趋势，我们使用周期和族的信息 [68], ，或者更准确地说，使用元素在 s区、p区、d区和f区中的位置，而不是原子序数。例如，所有卤素都会形成类似类型的键，而键长会随着周期升高而增加。

此外，我们提出，不确定性量化能够支持采用可调节的机器学习势参数。可调节参数要求实时探查机器学习势的可靠性，以避免对每次机器学习势训练扩展进行全面验证。只要机器学习势在生产计算中的误差低于容差，学习过程中的微小变化就不会显著影响结果。此外，具备不确定性量化的机器学习势还能在出现

未知构象时发出警告，即使这些构象位于一般训练空间内。因此，可以利用不确定性量化来估计机器学习势的可迁移性；不确定性越低，意味着可迁移性越高。例如，为了对 HDNNP 进行不确定性量化，一些研究采用了集成方法 [69–73]。随后，使用相同的参考数据，以不同方式独立训练一小组 HDNNP。由于神经网络具有高度灵活性，单个高维神经网络势的预测之间的偏差可用作预测误差的替代指标。此外，对集成取均值能够提高表征的准确率。我们指出，原子模拟中的任何建模方法都应配套采用此类不确定性度量[74], ，尽管相关实现直到最近才开始出现。

因此，本研究提出元素包容型原子中心对称函数（eeACSFs），以克服随元素数量增加而产生的不利扩展问题，并使其适用于任意数量的元素。结合不确定性量化和持续学习策略，我们提出了终身机器学习势（lMLP）的概念，它能够通过连续输入的新数据流以滚动方式进行训练。为实现这一具有挑战性的训练过程，我们提出了新型持续稳健（CoRe）优化器，这是一种用于随机一阶迭代优化的自适应方法。CoRe 与我们的终身训练策略结合使用，可自适应地选择所用训练数据，包括缩减数据集规模和剔除存疑数据。我们通过一个包含 42 个不同 S_N2 反应、涉及 10 种不同化学元素的数据集，展示了 lMLP 概念的性能。尽管我们关于 lMLP 概念的研究基于第二代高维神经网络势表示，但我们强调，该概念也可应用于其他基础模型。此外，CoRe 优化器和终身自适应数据选择也可用于训练 lMLP 以外的机器学习模型，并有助于推动终身机器学习的发展，以减轻灾难性遗忘。

本文的组织结构如下：第2节概述高维神经网络势方法，并介绍扩展等变原子中心对称函数、CoRe 优化器、终身自适应数据选择、不确定性量化和 lMLP。在第3节介绍计算细节之后，第4节首先描述参考数据，随后评估扩展等变原子中心对称函数表征的性能，将 CoRe 的结果与其他优化器的结果进行比较，测试终身训练策略的性能，并验证势能面预测中不确定性量化的有效性。最后，第5节给出结论。

2. 方法

2.1. 标准化的高维神经网络势

对于一个包含 N_{elem} 种元素且含有 N^m_{atom}个元素 m原子的体系，第二代高维神经网络势的能量 [17–19] 由每个原子的原子能贡献 $E^m_{\mathrm{atom},\ n}$ 求和得到 n:

$$E = \sum_{m=1}^{N_{\mathrm{elem}}} \sum_{n=1}^{N^m_{\mathrm{atom}}} E^m_{\mathrm{atom},n} \ . \tag{1}$$

每个原子能量贡献均由一个前馈神经网络计算得到,

$$E^m_{\mathrm{atom},n} = b^{m,3}_1 + \sum_{\lambda=1}^{n_2} a^{m,23}_{\lambda 1} \cdot f^2 \Bigg\{ b^{m,2}_\lambda + \sum_{\kappa=1}^{n_1} a^{m,12}_{\kappa\lambda} \cdot f^1 \left[b^{m,1}_\kappa + \sum_{i=1}^{n_G} a^{m,01}_{i\kappa} \cdot \alpha^m_i \left(G^m_{n,i} - \beta^m_i \right) \right] \Bigg\} \ . \tag{2}$$

权重 a 和 b 针对每种元素 m 分别进行训练，训练依据是包含多种化学结构的训练数据集中的能量和原子力。需要指出的是，隐藏层的数量可以不同于图示情况；图中网络包含两个隐藏层，分别有 n_1 和 n_2个神经元。将维度为 n_G的局域结构描述符向量$\mathbf{G}^m_n$作为该原子神经网络的输入，相关说明见下一节。我们提出了激活函数 $f(x) = 1.59223 \cdot \tanh(x)$，并在补充信息的 S1.1 节中对其进行讨论，同时介绍了专门设计的权重初始化方案 [40, 75]。与双曲正切函数和非专门设计的权重初始化相比，该激活函数和权重初始化方案在收敛优势和精度提升方面的表现见补充信息中的表S3以及图 S1 (a)、S1 (b) 和 S2。

与原始高维神经网络势方法相比， [17] 我们引入了可训练权重 α^m_i和 β^m_i，用于对结构描述符输入$G^m_{n,\ i}$进行标准化。为此，将结构描述符减去 β^m_i；该值初始化为初始训练数据集中所有属于相应元素 n 的原子 m的描述符值均值$G^m_{n,\ i}$。然后将结果乘以 α^m_i进行缩放；该值初始化为相应描述符值在初始训练数据中的标准差倒数。这样，输入层的权重 $a^{m,\ \ 01}_{\ i\kappa}$ 便乘以一组数值：对于初始训练数据，这些数值以零为中心，标准差为 1。这种标准化可以提升训练性能，并可根据新增训练数据进行调整。我们指出，可将权重初始化限制在特定区间内，以避免数值问题。权重对 $a^{m,\ \ 01}_{\ i\kappa}$ 和 α^m_i可以合并，但出于技术原因，我们在优化过程中将它们分开处理。

2.2. 元素包容型原子中心对称函数

高维神经网络势通常采用多体原子中心对称函数向量 [58] 来表示局部原子环境。这些描述符满足势能面的平移、旋转和置换不变性。它们取决于半径为 R_c的截断球内所有相邻原子之间的距离和夹角。该半径需要足够大，以涵盖所有相关相互作用。由于不使用连接关系，高维神经网络势能够描述化学反应。输入向量的维数不取决于具体的原子环境——这是构建普遍适用的原子神经网络所必需的条件。传统的 ACSF 向量采用这样的构造方式：每个原子中心对称函数仅表示特定化学元素对或三元组之间的所有相互作用。然而，这种构造会导致描述符数量随元素数量增加而不利地扩展。

因此，我们引入元素包容型原子中心对称函数（ eeACSF），它显式地依赖于来自元素周期表的元素信息 H （见下一段）。与原子中心对称函数类似，eeACSF 分为两类。径向 eeACSF，

$$G_{n,i}^{\mathrm{rad}} = \left[\frac{1}{H_{\mathrm{max},i}^{\mathrm{rad}}} \sum_{j \neq n}^{N_{\mathrm{atom}}} H_{i,j}^{\mathrm{rad}} \exp\left(-\eta_i^{\mathrm{rad}} \cdot R_{nj}^2\right) \cdot f_{\mathrm{c}}\left(R_{nj}\right) \right]^{\frac{1}{2}} , \tag{3}$$

是中心原子 n 与相邻原子 j之间距离 R_{nj} 的函数。角向 eeACSF，

$$G_{n,i}^{\mathrm{ang}} = \left\{ \frac{2^{-\zeta_i}}{H_{\mathrm{max},i}^{\mathrm{ang}}} \sum_{j \neq n}^{N_{\mathrm{atom}}} \sum_{k \neq n,j}^{N_{\mathrm{atom}}} H_{i,jk}^{\mathrm{ang}} \left[1 + \lambda_i \cos\left(\theta_{njk}\right)\right]^{\zeta_i} \cdot \exp\left[-\eta_i^{\mathrm{ang}} \left(R_{nj}^2 + R_{nk}^2\right)\right] \cdot f_{\mathrm{c}}\left(R_{nj}\right) \cdot f_{\mathrm{c}}\left(R_{nk}\right) \right\}^{\frac{1}{2}} , \tag{4}$$

除其他因素外，还取决于原子 n 与两个相邻原子 j 和 k 之间的夹角 θ_{njk}。对每个 eeACSF i 采用不同的参数值 $\eta_i^{\mathrm{rad}} \geq 0$、 $\eta_i^{\mathrm{ang}} \geq 0$、 $\lambda_i = \pm 1$和 $\xi_i \geq 1$ ，最终可生成原子 n周围原子环境的结构指纹。截断函数，

$$f_{\mathrm{c}}\left(R_{nj}\right) = \begin{cases} \exp\left[1 - \left(1 - \frac{R_{nj}^2}{R_{\mathrm{c}}^2}\right)^{-1}\right] & \text{for } R_{nj} < R_{\mathrm{c}} \\ 0 & \text{otherwise} \end{cases} , \tag{5}$$

使截断半径 R_{c} 之外的贡献平滑地衰减至零。这种截断函数的优点在于，在 R_{c} 处所有阶的导数也都为零，这有利于计算力、简正模式等。与原子中心对称函数（ACSFs）不同，扩展等变原子中心对称函数（eeACSFs）采用平方根，以减弱邻居数量对 eeACSF 值过强的影响。当训练数据包含分子尺寸和粒子密度差异很大的样本时，这种强烈影响就会显现，并可能因输入值的范围过宽而降低参数化性能。

作为元素描述符，我们提出采用 n 表示元素在元素周期表中的周期数， m表示 s 区和 p 区（主族 1 至 8）的族序数，以及 d 表示 d 区的族序数。主族元素满足 $d = 0$ ，而 d 区元素满足 $m = 0$。氦是一个特殊情况，其 $m = 8$。相邻原子 j 的这些描述符值用于径向 eeACSF 中与元素相关的项 $H_{i,\,j}^{\mathrm{rad}}$

$$H_{i,j}^{\mathrm{rad}} \in \{1,\ n_j,\ m_j,\ d_j,\ \overline{n}_j \coloneqq X - n_j, \\ \overline{m}_j \coloneqq 9 - m_j,\ \overline{d}_j \coloneqq 11 - d_j\} \ . \tag{6}$$

在构造不含元素依赖项的径向 eeACSF 时，可采用 $H_{i,\,j}^{\mathrm{rad}} = 1$ 。本研究将最大周期设为 $X - 1 = 5$，即考虑氩及之前的元素。采用描述符 n、 m 和 d，以平衡轻元素与重元素，以及价电子较少与较多的元素所作的贡献。对于主族元素，使用 $d = 0$ ；对于 d区元素，使用 $m = 0$。对于更高周期的元素，f区中的族可按与 d区相同的方式实现，并且 $X - 1$ 可设为 7。为使每种相互作用的贡献值处于 0 到 1 之间，径向 eeACSF 除以 $H_{\mathrm{max},i}^{\mathrm{rad}}$ ，该值是$H_{i,j}^{\mathrm{rad}}$ 的最大可能值。

角向 eeACSF 的元素依赖项 $H_{i,jk}^{\mathrm{ang}}$ 以线性组合的形式计算：

$$H_{i,jk}^{\mathrm{ang}} = \left|H_{i,j}^{\mathrm{rad}} + \gamma_i H_{i,k}^{\mathrm{rad}}\right| + C_i \ , \tag{7}$$

其中 $\gamma_i = \pm 1$ 和

$$C_i = \begin{cases} 0 & \text{for } \gamma_i = 1 \vee H_{i,j}^{\mathrm{rad}} = H_{i,k}^{\mathrm{rad}} = 0 \\ 1 & \text{otherwise} \end{cases} . \tag{8}$$

因此，角向元素嵌入原子中心对称函数的元素依赖前因子定义为

$$H_{\mathrm{max},i}^{\mathrm{ang}} = \begin{cases} 2H_{\mathrm{max},i}^{\mathrm{rad}} & \text{for } \gamma_i = 1 \\ H_{\mathrm{max},i}^{\mathrm{rad}} & \text{otherwise} \end{cases} . \tag{9}$$

总之，对于含有/不含有 d区元素的体系，分别需要五种/七种不同的 $H_{i,\,j}^{\mathrm{rad}}$ 以及九种/十一种 $H_{i,\,jk}^{\mathrm{ang}}$ 项。结合通常约五个 η_i^{rad}、两个 η_i^{ang} 、两个 λ_i和三个 ζ_i 参数值，描述符向量将由 25/35 个径向和 108/132 个角向 eeACSF 构成，且与元素数量无关。相比之下，径向原子中心对称函数的数量与元素数量成正比，而其数量

角向原子中心对称函数的数量则随 $\sum_{m=1}^{N_{\mathrm{elem}}} m$ 增长。例如，对于含有四种元素的情况，在参数值数量相同的条件下，可得到 20 个径向和 120 个角向原子中心对称函数。因此，元素数量约为四种时，计算成本达到平衡点，因为根据所列的依赖于元素的数值确定 $H_i^{\mathrm{rad}}{}_{,\,j}$ 和 $H_i^{\mathrm{ang}}{}_{,\,jk}$ 所需的额外工作量很小。我们注意到，不同参数值的数量会影响表示的分辨率。在机器学习势应用中，描述符计算的计算瓶颈在于根据原子位置求导，以获得原子力。与原子中心对称函数相比，扩展等变原子中心对称函数没有包含额外的依赖于位置的性质，因此每个扩展等变原子中心对称函数的计算成本与原子中心对称函数相近。

2.3. 高维神经网络势的训练

为使原子神经网络的权重在势能 $E^{\mathrm{ref}}{}_{,\,r}$ 以及参考构象 r的笛卡尔原子力分量 $F^{\mathrm{ref}}{}_{,\,\alpha\,n}{}^{r}{}_{,}$ 方面达到最优，定义一个损失函数：

$$L^t = \frac{q^2}{N_{\mathrm{conf}}} \sum_{r=1}^{N_{\mathrm{conf}}} \left(\frac{E^{r,t} - E^{\mathrm{ref},r}}{N_{\mathrm{atom}}^r} \right)^2 + \left(3 \sum_{r=1}^{N_{\mathrm{conf}}} N_{\mathrm{atom}}^r \right)^{-1} \cdot \sum_{r=1}^{N_{\mathrm{conf}}} \sum_{n=1}^{N_{\mathrm{atom}}^r} \sum_{\alpha=x,y,z} \left(F_{\alpha,n}^{r,t} - F_{\alpha,n}^{\mathrm{ref},r} \right)^2 . \tag{10}$$

原子力分量是能量相对于构象 r中原子 n 的笛卡尔坐标 α_n^r =x_n^r、 y_n^r、 z_n^r 的负梯度，

$$F_{\alpha,n}^r = -\frac{\partial E^r}{\partial \alpha_n^r} . \tag{11}$$

每个训练轮次 t 中同时训练一组 N_{conf} 构象，以加速优化并减少单个数据点的过拟合。由于 HDNNP 对某个构象的预测可能取决于不同化学元素对应的原子神经网络，因此这些网络也会同时训练。为平衡能量和力对损失函数的贡献，使用超参数 q 。该超参数需要谨慎选择，因为它会显著影响训练性能。为优化权重，在每个训练轮次中，计算损失函数相对于权重的梯度，

$$w_\xi \coloneqq w_{\chi,\kappa\lambda}^{m,\mu\nu} \in \{ a_{\kappa\lambda}^{m,\mu\nu},\ b_\lambda^{m,\nu},\ \alpha_\kappa^m,\ \beta_\kappa^m \} , \tag{12}$$

计算得出，其中 ξ 是每个权重的唯一索引。

2.4. 持续韧性（CoRe）优化器

为了改进训练过程，我们开发了持续弹性（CoRe）优化器，旨在结合弹性反向传播（RPROP）[76, 77]的鲁棒性与Adam优化器 [78]的性能。此外，我们引入了损失函数梯度移动平均的自适应衰减率、根据重要性评分得到的权重可塑性因子，以及用于限制权重值的权重衰减，从而使收敛速度和最终准确率超越最先进的优化器。

CoRe 是一种一阶梯度优化器。它根据优化历史，为每个权重w_ξ采用单独的自适应学习率。该算法适用于随机迭代优化，即在每个训练轮次中使用训练数据批次的子样本。因此，可以利用随机梯度下降（SGD） [79]的计算效率优势。

遵循 Adam 优化器的思路，梯度的指数移动平均值

$$g_\xi^\tau = \beta_1^\tau \cdot g_\xi^{\tau-1} + (1 - \beta_1^\tau) \frac{\partial L^t}{\partial w_\xi^t} , \tag{13}$$

以及平方梯度的指数移动平均值

$$h_\xi^\tau = \beta_2 \cdot h_\xi^{\tau-1} + (1 - \beta_2) \left(\frac{\partial L^t}{\partial w_\xi^t} \right)^2 , \tag{14}$$

在计算各权重 w_ξ^t的自适应学习率时，会启用衰减率为 β_1^τ, $\beta_2 \in [0, 1)$ 的相关机制。 τ 是每个权重的优化步计数器，其数值可能与训练轮次t不同。例如，对于 HDNNP，每个训练轮次中的训练数据子样本可能不包含整个训练数据集中的所有化学元素，因此在相应的训练轮次中，只会更新部分原子神经网络的权重。这样一来，HDNNP 模型本身便提供了一种架构策略，以减轻灾难性遗忘。需要指出的是，公式（13）和公式（14）表示最小化的情况；若要进行最大化，则必须反转损失函数对权重的导数的符号。

与 Adam优化器不同， β_1 是τ的函数，

$$\beta_1^\tau = \beta_1^{\mathrm{b}} + \left(\beta_1^{\mathrm{a}} - \beta_1^{\mathrm{b}} \right) \exp \left[- \left(\frac{\tau - 1}{\beta_1^{\mathrm{c}}} \right)^2 \right] . \tag{15}$$

超参数 $\beta_1^{\mathrm{a}}, \beta_1^{\mathrm{b}} \in [0, 1)$ 定义了 β_1 的初始值和最终值。这些值通过一个超参数为 $\beta_1^{\mathrm{c}} > 0$ 的高斯函数相互转换。较大的 β_1 值会增加对先前梯度信息的依赖，也就是对所有近期训练数据子样本的优化历史的依赖。因此，可以提高相对于整个训练数据集的优化性能

。较小的 β_1 值会使其更加依赖当前子样本的梯度，从而在一定程度上降低优化过程的转动惯量。这有利于快速收敛初期出现的快速且剧烈变化的梯度；此时优化过程从随机初始化的权重开始。总而言之，可在优化过程中利用公式（15）增大 β_1 ，以提升训练性能。

CoRe 权重更新幅度的一项贡献，类比 Adam优化器，按以下方式得到

$$u_\xi^\tau = \frac{g_\xi^\tau}{1-(\beta_1^\tau)^\tau}\left\{\left[\frac{h_\xi^\tau}{1-(\beta_2)^\tau}\right]^{\frac{1}{2}}+\epsilon\right\}^{-1} . \tag{16}$$

移动平均 g_ξ^τ 和 h_ξ^τ 分别通过$1-(\beta_1^\tau)^\tau$和 $1-(\beta_2)^\tau$进行偏差校正，以抵消它们趋向于零的初始化偏差 (g_ξ^0, $h_\xi^0 = 0$)。将两个经过偏差校正的移动平均相除，可使权重更新幅度不受梯度缩放的影响。由此实现了一种步长退火，因为梯度值减小会导致更新 u_ξ^τ减小。因此，对于表现良好的优化， u_ξ^τ 的绝对值通常会从优化第一步 $\tau = 1$ 的 ± 1 逐渐趋近于零。超参数 β_2 的值越大，这种退火效果越明显。加入超参数 $\epsilon \gtrapprox 0$ 以确保数值稳定性。

作为对权重更新的第二项贡献，可塑性因子

$$P_\xi^\tau = \begin{cases} 0 & \text{for } \tau > t_{\text{hist}} \\ & \wedge\ S_\xi^{\tau-1} \text{ top-}n^{m,\mu\nu}_{\text{frozen},\chi} \text{ in } \mathbf{S}^{m,\mu\nu,\tau-1}_\chi \\ 1 & \text{otherwise} \end{cases} , \tag{17}$$

用于减轻旧信息的遗忘。在初始训练阶段，该因子等于1。当$\tau > t_{\text{hist}}$时，通过超参数 $t_{\text{hist}} > 0$，可塑性因子可以将P_ξ^τ 设为零，从而冻结部分权重的值。这些权重的选择取决于评分值 $S_\xi^{\tau-1}$ ，该评分值根据先前预测的损失函数下降量对权重的重要性进行排序，具体将在公式（20）中说明。评分值 $S_\xi^{\tau-1}$ 会与同一族中的所有其他评分值$\mathbf{S}^{m\ \mu\nu\ \tau-1}_{\chi\ ,,}$进行比较。对于 HDNNP，同一族由同一元素 m的原子神经网络中具有相同权重类型χ和相同层分配 $\mu\nu$的权重组成。各自所属族中评分值最高的 n_{frozen} 个权重会在优化步骤 τ中被冻结。超参数 $n^{m,\mu\nu\ \text{已冻结}}_{\chi} \geq 0$ 可以针对每个权重族单独设置。这样便对估计重要性最高的权重实施了正则化。

步长调整的另一项贡献是借鉴自 RPROP，

$$s_\xi^\tau = \begin{cases} \min\left(\eta_+ \cdot s_\xi^{\tau-1}, s_{\max}\right) & \text{for } g_\xi^{\tau-1}\cdot g_\xi^\tau \cdot P_\xi^\tau > 0 \\ \max\left(\eta_- \cdot s_\xi^{\tau-1}, s_{\min}\right) & \text{for } g_\xi^{\tau-1}\cdot g_\xi^\tau \cdot P_\xi^\tau < 0 \\ s_\xi^{\tau-1} & \text{for } g_\xi^{\tau-1}\cdot g_\xi^\tau \cdot P_\xi^\tau = 0 \end{cases} . \tag{18}$$

一般而言，步长 s_ξ^τ 不取决于梯度的大小，而只取决于其符号，因此能够实现稳健的优化。每当 g_ξ^τ 相较于$g_\xi^{\tau-1}$ 改变符号时，前一步优化很可能越过了某个局部极小值。这意味着前一步的步长 $s_\xi^{\tau-1}$ 过大，需要减小。如果 g_ξ^τ和 $g_\xi^{\tau-1}$ 的符号相同，则可以增大$s_\xi^{\tau-1}$ 以加快收敛。减小因子 η_- 和增大因子η_+，结合 $0 < \eta_- \leq 1 \leq \eta_+$，用于得到当前步长 s_ξ^τ，其中 s_ξ^τ 受最小和最大步长 $s_{\min}$， $s_{\max} > 0$ 的限制。如果 P_ξ^τ、 g_ξ^τ和/或 $g_\xi^{\tau-1}$ 为零，则省略步长更新。需要初始化步长 s_ξ^0 ，以确定第一个优化步长 s_ξ^1。该值可根据初始权重值按合理比例选取。经验表明，由于适应速度很快，该参数的精确取值并不十分重要。我们注意到，对于 $g_\xi^{\tau-1}\cdot g_\xi^\tau < 0$，梯度不会重置为 $g_\xi^\tau = 0$；这与某些包含回溯权重步长 [76, 80]的RPROP变体不同。因此，可以保留 g_ξ^τ 的历史信息。

为降低过拟合风险，权重更新

$$w_\xi^t = \left(1 - d_\chi^{m,\mu\nu}\cdot\left|u_\xi^\tau\right|\cdot P_\xi^\tau\cdot s_\xi^\tau\right) w_\xi^{t-1} - u_\xi^\tau \cdot P_\xi^\tau \cdot s_\xi^\tau , \tag{19}$$

包含权重衰减，其权重衰减超参数为 $d^{m}_{,\chi\ \mu\nu} \in [0, (s_{\max})^{-1})$。权重 w_ξ^{t-1}会按由 $d^{m}_{,\chi\ \mu\nu}$与当前权重更新的绝对值 $|u_\xi^\tau|\cdot P_\xi^\tau \cdot s_\xi^\tau$的乘积所得的比例进行缩减。需要注意的是，权重更新的符号（与 g_ξ^τ的符号相同）仅编码在 u_ξ^τ 中。权重衰减超参数的倒数是表现良好的优化中的最大权重绝对值，即 $u_\xi^\tau \leq \pm 1$，从而防止权重大幅增加或减少。为获得新权重 w_ξ^t，需从先前权重 w_ξ^{t-1} 中减去当前权重更新 $u_\xi^\tau \cdot P_\xi^\tau \cdot s_\xi^\tau$。因此，先前权重会沿着与 g_ξ^τ 符号相反的方向，按照单独调整的学习率发生变化。

评分值，

$$S_\xi^\tau = \begin{cases} S_\xi^{\tau-1} + (t_{\text{hist}})^{-1}\, g_\xi^\tau \cdot u_\xi^\tau \cdot P_\xi^\tau \cdot s_\xi^\tau & \text{for } \tau \leq t_{\text{hist}} \\ \left[1-(t_{\text{hist}})^{-1}\right] S_\xi^{\tau-1} & \\ \quad + (t_{\text{hist}})^{-1}\, g_\xi^\tau \cdot u_\xi^\tau \cdot P_\xi^\tau \cdot s_\xi^\tau & \text{otherwise} \end{cases} , \tag{20}$$

用于衡量权重对先前损失函数下降的特定贡献。该方法受突触智能方法 [53] 的启发。损失函数下降估计为

每一步中梯度 g_ξ^τ 的移动平均与权重更新 $u_\xi^\tau \cdot P_\xi^\tau \cdot s_\xi^\tau$ 的乘积 τ。对于沿梯度反方向的无穷小变化，梯度与变化量的乘积等于相应的损失函数下降。对于本例优化中的较大更新，该乘积通常会高估损失函数下降，但仍是合理的估计。此外，如果梯度是根据整个训练数据集的子样本计算的，就会带有噪声。在评分值的计算中，省略了更新的符号反转。因此，较高的正评分值表示损失函数下降幅度较大。评分值初始化为 $S_\xi^0 = 0$。对于每个 $\tau \leq t_{\text{hist}}$，将 $g_\xi^\tau \cdot u_\xi^\tau \cdot P_\xi^\tau \cdot s_\xi^\tau$ 以相等的贡献（t_{hist}）$^{-1}$ 累加，以建立初始历史记录。之后，S_ξ^τ 以衰减参数为 $1 -$（t_{hist}）$^{-1}$ 的指数移动平均方式计算。评分值用于识别对准确预测先前训练数据最重要的权重。随后可通过可塑性因子（公式 (17)）对这些权重加以限制，以平衡稳定性-可塑性比率。

作为一种变体，CoRe 还可以使用梯度移动平均值的符号 g_ξ^τ 作为更新因子 u_ξ^τ，即 u_ξ^τ =sgn(g_ξ^τ)。此外，CoRe 也可以在不使用可塑性因子的情况下运行，即 $n^{m,\ \mu\nu\ \text{冻结}\ \chi} = 0$。对于超参数设置$\beta_1^{\text{a}} = \beta_1^{\text{b}}$, $n^{m,\ \mu\nu\ \text{冻结}\ \chi} = 0$, η_+, $\eta_- = 1$和 $d^{m,\ \chi\ \mu\nu} = 0$，Adam优化器是 CoRe 的一种特殊情况。不使用回溯权重步长的 RPROP 则是另一种特殊情况，其超参数设置为 β_1^{a}, β_1^{b}, β_2 =0, $n^{m,\ \mu\nu\ \text{冻结}\ \chi} = 0$ 和 $d^{m,\ \chi\ \mu\nu} = 0$。

可针对 $\beta_2 = 0.999$、$\epsilon = 10^{-8}$、$\eta_- = 0.5$、$\eta_+ = 1.2$、$s_{\min} = 10^{-6}$和 $s_{\max} = 1$提供超参数取值的一般建议，这些建议大多与 Adam优化器和 RPROP 一致。通常，$s_\xi^0 = 10^{-3}$ 是初始学习率的良好选择。权重衰减 $d^{m,\ \chi\ \mu\nu}$取决于期望的权重值范围。例如，原子神经网络中与输出神经元相关联的权重不应受到限制，以允许原子能取任意值；而连接输入层与隐藏层或连接隐藏层与隐藏层、且以双曲正切函数作为激活函数的权重，则可通过 $d^{m,\ \chi\ \mu\nu} = 0.1$加以限制。$\beta_1^{\text{c}}$ 和 t_{hist} 取决于收敛速度和训练轮次总数。例如，本研究采用了 $\beta_1^{\text{c}} = 500$ 和 $t_{\text{hist}} = 500$，训练轮次总数则为 1500、2000年和 2500。此外，$n^{m,\ \mu\nu\ \text{冻结}\ \chi}$ 设为权重数量的 1% $a^{m,\ \mu\nu}_{\kappa\lambda}$，并分别在各自的族中设为 $b^{m,\ \nu}_{\ \lambda}$，与输出神经元相关联的权重除外。对于变化迅速且幅度较大的梯度，β_1^{a} 和 β_1^{b} 需要设置得更小。例如，本研究使用了 $\beta_1^{\text{a}} = 0.45$和 $\beta_1^{\text{b}} = 0.7$，0.725。我们注意到，对于相关性很低的训练数据，增大子样本大小有助于减少梯度波动。

2.5. 终身自适应数据选择

随机优化，即在每个训练轮次中基于训练数据的子样本计算梯度，可以将大型数据集的计算需求控制在可承受的水平。然而，随机抽取子样本可能导致方法效率低下，因为不同数据通常具有不同的准确率水平。如果在终身机器学习的训练过程中加入更多训练数据，这一问题会更加严重。此外，在训练过程中需要移除冗余数据和错误数据，才能自主地从连续的新数据流中学习。尽管如此，回放有代表性的旧数据对于避免灾难性遗忘仍然十分重要。

为了解决这些问题，我们开发了一种用于终身自适应数据选择的算法。该算法的第一个关键要素是为每个训练构象 r 设置一个自适应选择因子 S_{hist}^r，该因子取决于该构象对损失函数的贡献以及训练历史。第二个关键要素是为随机选择优质和表示不足的训练数据制定平衡方案。该算法旨在排除冗余和错误数据，并通过平衡对新数据或表示不足数据的调整与对优质数据的回放，提升训练表现。

算法1： 选择要拟合的训练数据子样本
，以便在终身自适应数据
选择算法的一个训练轮次中进行拟合。所有向量运算均逐元素进行。
包含条件的赋值语句仅更改
条件为真的向量条目。

$N_{\text{fit}} \leftarrow$ 最小值 $[N_{\text{fit}}$,长度 $(\mathbf{S}^{>0}$ 历史$)]$
$N_{\text{good}} \leftarrow$ 整数（$p_{\text{good}} \cdot N_{\text{fit}}$）
$N_{\text{bad}} \leftarrow N_{\text{fit}} - N_{\text{good}}$
$L_{\text{old}}^{\max} \leftarrow$ 最大 $\left(\mathbf{L}^{\backslash \text{fit}}_{\text{旧}}\right)$
$\mathbf{P}_{\text{bad}} \leftarrow \mathbf{S}_{\text{hist}}\ \mathbf{L}^{\text{old}} / L_{\text{old}}^{\max}$
旧$\mathbf{P}_{\text{bad}} \leftarrow$ 最大（$\mathbf{S}_{\text{hist}}$）如果 $P_{\text{bad}}^r =$ 不是数字
$\mathbf{P}_{\text{bad}} \leftarrow \mathbf{P}_{\text{bad}} /$ 总和 $(\mathbf{P}_{\text{bad}})$
$\mathbf{D}_{\text{fit}} \leftarrow$ 随机_选择 $(\mathbf{D}^{\backslash \text{fit}}_{\text{旧}}, \mathbf{P}_{\text{bad}}, N_{\text{bad}})$
$\mathbf{P}_{\text{good}} \leftarrow \mathbf{S}^{>0}_{\text{历史}} \left[\mathbf{L}^{\text{旧}} / L_{\text{old}}^{\max}\right] - 1) \cdot \epsilon'$
$P_{\text{good}}^{\min} \leftarrow$ 分钟 $(\mathbf{P}_{\text{g}}^{>0}$旧
$\mathbf{P}_{\text{good}} \leftarrow P_{\text{good}}^{\min}$ 如果 $L^{\backslash \text{fit}}_{\text{旧}\ r} = L_{\text{old}}^{\max}$
$\mathbf{P}_{\text{good}} \leftarrow P_{\text{good}}^{\min}$ 如果 $P_{\text{good}}^r =$ 非数字
$\mathbf{P}_{\text{good}} \leftarrow \mathbf{P}_{\text{good}} /$ 总和 $(\mathbf{P}_{\text{good}})$
$\mathbf{D}_{\text{fit}} \leftarrow \mathbf{D}_{\text{fit}} \cup$ 随机_选择 $\left(\mathbf{D}^{\backslash \text{fit}}, \mathbf{P}_{\text{good}}, N_{\text{good}}\right)$

终身自适应数据选择算法可分为两部分：一部分在计算损失函数之前执行，另一部分在计算之后执行。在第一部分（算法1）中，选择用于计算损失函数及相应梯度的训练数据子样本。该子样本标记为索引“fit”，因为这些数据用于拟合权重。每个训练轮次中要拟合的构象数量

N_{fit}是一个训练超参数，可根据数据集大小以及数据点之间的相关性来选择。由于自适应选择因子 S^r_{hist}（算法 2）可能会排除某些构象，因此 N_{fit} 不能大于满足 $S^r_{\text{hist}} > 0$的构象数量，即 $\text{len}\left(\mathbf{S}^{>0}_{\text{hist}}\right)$。训练数据子样本分为两组：一组包含 N_{good} 个已得到充分表示的构象，另一组包含 N_{bad} 个新的或表示不足的构象。优质数据的比例由 p_{good}定义（算法 2），而 N_{good} 必须为整数。p_{good} 初始化为 0。

为了确定构象分别被选为较差或较好数据的概率 $\mathbf{P}_{\text{bad}}$ 和$\mathbf{P}^{\text{good}}$，采用构象对损失函数的贡献。最后计算出的贡献$\mathbf{L}_{\text{old}}$（如果此前尚未计算某个构象的贡献，则为非数值（NaN））除以最大的非 NaN 贡献$\mathbf{L}^{\max}_{\text{old}}$。这些商再乘以构象特定的自适应选择因子$\mathbf{S}_{\text{hist}}$，从而得到概率$\mathbf{P}_{\text{bad}}$。$\mathbf{S}_{\text{hist}}$ 初始化为单位向量。将$\mathbf{P}_{\text{bad}}$ 的所有 NaN 分量设为$\mathbf{S}_{\text{hist}}$中的最大值。为进行归一化，将$\mathbf{P}_{\text{bad}}$ 除以其分量之和。从训练数据集$\mathbf{D}$ 中按概率$\mathbf{P}_{\text{bad}}$ 随机选择 N_{bad} 个构象，作为训练数据子样本的一部分。

为了从剩余数据集$\mathbf{D}^{\setminus\text{fit}}$中获得 N_{good} 较好数据，需确定概率$\mathbf{P}^{\text{good}}$。$\mathbf{P}^{\text{good}}$ 与单位向量$\mathbf{1}$和$\mathbf{L}^{\setminus\text{fit}}_{\text{old}}$ 与 $L^{\max}_{\text{old}}$ 的商之间的差成正比。确定$\mathbf{P}^{\text{good}}$ 中最小的非零概率，并将其乘以 $\epsilon' \gtrsim 0$，以获得一个很低的概率 $P^{\min}_{\text{good}}$，用于未被排除但不受青睐的构象。如果相应构象的损失函数贡献等于最大贡献，或者$\mathbf{P}^{\text{good}}$ 对应项为 NaN，则将$\mathbf{P}^{\text{good}}$ 对应项设为 $P^{\min}_{\text{good}}$。随后，将$\mathbf{P}^{\text{good}}$ 除以其各项之和。较好数据子样本是从$\mathbf{D}^{\setminus\text{fit}}$中按概率$\mathbf{P}^{\text{good}}$随机选择的 N_{good} 个数据。然后，将选出的较差数据和较好数据合并，作为训练数据子样本来计算损失函数。

为更新自适应选择因子 $\mathbf{S}_{\text{hist}}$，使用训练数据子样本中各构象的新损失函数贡献$\mathbf{L}^{\text{fit}}_{\text{new}}$（算法 2）。因此，采用能量损失$\mathbf{L}^{\text{fit}}_{E}$ 和力损失 $\mathbf{L}^{\text{fit}}_{F}$ 的加权和，并除以相应构象的原子数 $\mathbf{N}^{\text{fit}}_{\text{atom}}$。$\mathbf{L}_{\text{rel}}$是$\mathbf{L}^{\text{fit}}_{\text{new}}$ 与训练轮次总损失 $L^{\text{fit}}_{\text{new}}$ 的商，后者由公式（10）定义。将该商与不同阈值 $T^1_F < 1 < T^2_F < T^3_F < T_{\mathbf{X}}$进行比较，以更新$\mathbf{S}_{\text{hist}}$ 和排除连续计数器 $\mathbf{X}$。引入后者是为了排除预测异常值。假设模型合理，异常值很可能是错误数据点。因此，$\mathbf{X}$（初始化为零向量）用于计数连续评估 L^r_{rel} 大于阈值 $T_{\mathbf{X}}$的次数。如果 X^r 达到超参数 $N_{\mathbf{X}}$ 时，自适应选择因子 S^r_{hist} 被设为零，因此相应构象将不再参与后续训练。

算法 2： 终身自适应数据选择算法中，更新训练数据子样本内每个构象的自适应选择因子。所有向量运算均按元素进行。包含条件的赋值语句仅更改满足条件的向量条目。

$\mathbf{L}^{\text{fit}}_{\text{new}} \leftarrow \dfrac{q^2\mathbf{L}^{\text{fit}}_{E} + \frac{1}{3}\mathbf{L}^{\text{fit}}_{F}}{\mathbf{N}^{\text{fit}}_{\text{atom}}}$
$\mathbf{L}_{\text{rel}} \leftarrow \dfrac{\mathbf{L}^{\text{fit}}_{\text{new}}}{\overline{L}^{\text{fit}}_{\text{new}}}$
$\mathbf{X} \leftarrow X^r + 1$ 如果 $L^r_{\text{rel}} > T_{\mathbf{X}}$
$\mathbf{X} \leftarrow 0$ 如果 $L^r_{\text{rel}} \le T_{\mathbf{X}}$
$\mathbf{S}_{\text{hist}} \leftarrow$最大值 $(1,\ S^r_{\text{hist}})$ 如果 $L^r_{\text{rel}} \ge T^1_F$
$\mathbf{S}_{\text{hist}} \leftarrow$最小值（$S^r_{\text{hist}}, 1$），如果 $L^r_{\text{rel}} \le T^2_F\mathbf{S}$
$_{\text{hist}} \leftarrow S^r_{\text{hist}} \cdot F_{--}$ 如果 $L^r_{\text{rel}} < T^1_F \wedge L^r_{\text{new}} \le L^r_{\text{old}}$
$\mathbf{S}_{\text{hist}} \leftarrow S^r_{\text{hist}} \cdot F_{-}$ 若 $L^r_{\text{rel}} < T^1_F \wedge L^r_{\text{new}} > L^r_{\text{old}}$
$\mathbf{S}_{\text{hist}} \leftarrow S^r_{\text{hist}} \cdot F_{+}$ 若 $T^2_F < L^r_{\text{rel}} \le T^3_F \wedge L^r_{\text{new}} > L^r_{\text{old}}$
$\mathbf{S}_{\text{hist}} \leftarrow S^r_{\text{hist}} \cdot F_{++}$ 若 $L^r_{\text{rel}} > T^3_F \wedge L^r_{\text{new}} > L^r_{\text{old}}$
$\mathbf{S}_{\text{hist}} \leftarrow 0$ 若 $X^r \ge N_{\mathbf{X}}$
$\mathbf{S}_{\text{hist}} \leftarrow 0$ 若 $S^r_{\text{hist}} < S^{\min}_{\text{hist}} \vee S^r_{\text{hist}} > S^{\max}_{\text{hist}} p_{\text{good}} \leftarrow$剪裁 $\left[p_{\text{good}} + \text{符号函数}\left(L^{\text{fit}}_{\text{新}} - \overline{L}^{\text{fit}}_{\text{旧}}\right)\overline{p}_{\pm},\ 0,\quad p^{\max}_{\text{good}}\right]$
$\mathbf{L}^{\text{fit}}_{\text{old}} \leftarrow \mathbf{L}^{\text{fit}}_{\text{new}}$
$\overline{L}^{\text{fit}}_{\text{旧}} \leftarrow \overline{L}^{\text{fit}}_{\text{new}}$

自适应选择因子 $\mathbf{S}_{\text{hist}}$ 还会根据以下条件进行进一步调整：如果 $L^r_{\text{rel}} \ge T^1_F$ 且 $S^r_{\text{hist}} < 1$，则将 S^r_{hist} 设为 1。如果 $L^r_{\text{rel}} \le T^2_F$ 且$S^r_{\text{hist}} > 1$，则也将 S^r_{hist} 改为 1。这样，充分表示的数据（$L^r_{\text{rel}} < T^1_F$）的 S^r_{hist}可以低于 1，从而降低其被选入训练数据子样本的概率。另一方面，表示不足的数据（$L^r_{\text{rel}} > T^2_F$）的 S^r_{hist}可以大于 1，从而提高其被选中的概率。对于 $T^1_F \le L^r_{\text{rel}} \le T^2_F$的构象，$S^r_{\text{hist}}$ 会重置为 1，因为L^r_{rel} 接近平均损失贡献。此外，$\mathbf{S}_{\text{hist}}$ 可以通过递减因子 F_{--} 和 F_{-} 以及递增因子 F_{+} 和 F_{++}进行调整。这样，S^r_{hist} 会跟踪训练历史，以评估相关构象 r的训练重要性。递减因子和递增因子分别初始化为 $F_{-(-)} = (S_{\min})^{(N_{F_{-(-)}})^{-1}}$ 和$F_{+(+)} = (S_{\max})^{(N_{F_{+(+)}})^{-1}}$。因此，超参数 $N_{F_{--}}$、$N_{F_{-}}$、$N_{F_{+}}$和 $N_{F_{++}}$ 决定了连续应用相应递减因子和递增因子的次数；达到这些次数后，自适应选择因子将分别低于和高于超参数 $S_{\min}$ 和$S_{\max}$。对于充分表示的数据，如果损失函数贡献保持不变或较先前计算值有所降低，则 S^r_{hist} 会降低 F_{--}。否则，充分表示的数据的 S^r_{hist} 会以较小幅度 F_{-} 降低。对于 $T^2_F \le L^r_{\text{rel}} \le T^3_F$ 和$L^r_{\text{new}} > L^r_{\text{old}}$，$S^r_{\text{hist}}$ 会增加 F_{+}；而对于 $L^r_{\text{rel}} \ge T^3_F$ 和 $L^r_{\text{new}} > L^r_{\text{old}}$，则会以较大的因子 F_{++} 增加。

随后，对于数据集合缩减，所有低于 S^r_{hist} $S_{\min}$ 的值都设为零

因为模型能够持续良好地表示相关构象。因此，这些数据可能是冗余的，可以将其排除，因为对其他数据进行复习和/或采用其他学习策略，足以避免遗忘这些构象。所有高于 S^{r}_{hist} $S_{\max}$ 的值也都设为零，因为经过许多优化步数后，模型仍无法准确表示这些构象。因此，假设模型合理，这些构象与其他大部分数据不一致，可以从训练数据中移除。这样一来，其他数据的训练性能便会提升。

为了自适应平衡对表示不足的数据的学习和不遗忘对表示良好的数据，良好数据的比例 p_{good} 会根据总损失的变化 $L^{\mathrm{fit}}{}_{\text{新}} - L^{\mathrm{fit}}{}_{\text{旧}}$进行调整。 p_{good} 若损失变化为正，则增加$p_{\pm}$；若损失变化为负，则减少$p_{\pm}$。 L_{old} 初始化为无穷大， $p_{\pm}$ 定义为 $\overline{p}^{\max}_{\mathrm{good}} \cdot (N_p)^{-1}$，其中最大 p_{good} 值为 $p^{\max}_{\mathrm{good}}$，可能的 p_{good} 值总数为 N_p。最后，训练数据子样本中各构象的新损失函数贡献 $\mathbf{L}^{\mathrm{fit}}_{\mathrm{new}}$ 会覆盖旧损失函数贡献 $\mathbf{L}^{\mathrm{fit}}_{\mathrm{old}}$，而新的总损失值 $L^{\mathrm{fit}}{}_{\text{新}}$ 会替代旧的总损失值 $L^{\mathrm{fit}}{}_{\text{旧}}$。

2.6. 不确定性量化

对于一种化学结构，lMLP 除了需要预测能量 ΔE 和原子力分量$\Delta F_{\alpha,\ n}$ 及其数值外，还需要预测它们的不确定性。原因在于，lMLP 的一个先决条件是能够在每个训练阶段都产生结果。由于不同训练阶段的预测可能有所不同，lMLP 需要提供预测准确度的度量。这样，便可以根据预测的准确度进行比较，并要求它们彼此一致。由此可获得一种可靠的方法，同时还能随着时间推移适应新信息。值得注意的是，不确定性量化也显著推动了其他计算化学方法的发展，因为这样可以避免过度解读超出方法准确度范围的特征。

机器学习模型的不确定性可分解为参考数据中的噪声、模型偏差和模型方差所造成的贡献[81]。含噪训练数据的处理由终身自适应数据选择解决，而测试数据中的噪声不会影响模型性能，只会影响性能评估。对于正确收敛的电子结构计算，噪声由收敛阈值决定，因此 MLP 参考数据中的噪声较小。模型偏差源于模型架构、描述符表示方式以及参考数据集规模的局限性。因此，需要针对不同应用优化模型架构，并测试更复杂的描述符，以便将这一误差贡献控制在较小水平。终身学习解决参考数据不完整所带来的剩余问题。通过集成（或委员会）模型可以可靠地降低模型方差误差，其大小也可以通过统计方法预测。相应的不确定性估计基于机器学习模型（例如人工神经网络）所依赖的数学表达式具有极大的灵活性这一特点。对于训练充分的构象空间，集成成员的预测通常会一致；而对于未知的构象空间，预测则是任意的。因此，由独立训练且训练方式各异的机器学习势组成的集成，其预测可能相差很大，例如使用不同的初始权重。

因此，研究人员使用机器学习势的集成来直接量化模型方差的贡献 [69–73, 82]。最终的能量预测 E 由集成中 N_{MLP}个机器学习势的能量均值 E_p得到，

$$\overline{E} = \frac{1}{N_{\mathrm{MLP}}} \sum_{p=1}^{N_{\mathrm{MLP}}} E_p \ . \tag{21}$$

类似地，原子力分量 $F_{\alpha,\ n}$ 通过计算机器学习势集成的负能量梯度相对于笛卡尔坐标 α_n 的均值来获得，

$$\overline{F}_{\alpha,n} = -\frac{1}{N_{\mathrm{MLP}}} \sum_{p=1}^{N_{\mathrm{MLP}}} \frac{\partial E_p}{\partial \alpha_n} \ . \tag{22}$$

随后可通过集成成员能量预测的样本标准差，得到能量不确定性 ΔE。

$$\Delta\overline{E} = \max\left\{ \mathrm{RMSE}(\overline{E}), \left[\frac{c^2}{N_{\mathrm{MLP}} - 1} \sum_{p=1}^{N_{\mathrm{MLP}}} \left(E_p - \overline{E}\right)^2 \right]^{\frac{1}{2}} \right\} , \tag{23}$$

而最小不确定性定义为参考数据的均方根误差 RMSE(E)。原子力分量的不确定性 $\Delta F_{\alpha,\ n}$ 也可以用类似的方法计算。

引入缩放因子 c 来调整样本标准差，从而提供一定的置信区间，使不确定性量化结果相对于参考数据大于或等于实际误差。然而，这种不确定性量化未考虑模型偏差的影响，因此，当训练数据不足是误差的主要来源时，会低估误差。尽管如此，由于这些情况下模型方差也会增大，即使预测的不确定性数值不准确，这种不确定性量化仍能指示误差是否会很大。对于机器学习势的应用，这些信息已足够，因为小的不确定性可以得到准确预测，而大误差的出现无论如何都需要改进

计算，而这可以通过终身机器学习实现。

2.7. 终身机器学习势的展望

从长远来看，宏伟目标是构建一种 lMLP，使其能够随着时间推移不断学习越来越多的体系，同时训练也能受益于先前学到的化学相互作用知识。这样，lMLP 的学习方式便与化学的发展方式一致，即通过系统地扩展已有知识来学习。由于本研究属于概念验证，所提出的终身训练策略自然尚未发展到足以实现这一宏伟目标的成熟阶段。从短期来看，采用本文提出的算法时，lMLP 大概会使用相关体系的数据进行训练，因为在这种情况下收益最大、难度最低。随着终身训练策略和机器学习势模型日益完善，可以将越来越多的信息整合到一个 lMLP 中。

本研究为 lMLP 奠定了全面的基础。未来工作需要开发相关算法，例如持续扩展模型架构，以避免底层深度学习表征的容量限制。例如，可以通过采用尺寸各异的 lMLP 集成成员架构来实现神经网络增长。通过跟踪它们的相对性能，可以确定是否需要扩展，并调整各个集成成员。此外，可以应用粗粒化结构标识符（如分箱与哈希方法）[83]，以补充用于回放的训练数据的筛选与排除标准。此外，这些标识符还可用于快速检查训练数据中是否有相似构象能够表征某一特定构象。根据这些信息，可以区分未知构象和表征不足的构象，从而避免冗余的参考计算。

当前机器学习势模型通常仅限于表示单一电子态。因此，例如，单个多层感知机通常只能表示特定的总电荷和自旋多重度。然而，近期研究致力于克服这些限制，以构建更通用的机器学习势模型[28, 33, 84]，进而可将其用作 lMLP 的基础模型。当前机器学习势的一个限制是要求训练数据保持一致，即所有训练数据都必须采用相同的方法、基组等进行计算。因此，需要开发机器学习势模型，例如，将不同参考方法视为不同电子态并以类似方式处理，从而建立可利用不同参考方法所得数据进行学习的 lMLP。

此外，公开、免费且易于获取的已发表研究参考数据显然至关重要。为了继续先前研究的训练，除了最终权重值外，还需要保存并提供权重 τ, g_ξ^τ, h_ξ^τ, s_ξ^τ, 和 S_ξ^τ 的相关 CoRe 优化器性质，以及每个仍需使用的训练构象的终身自适应数据选择性质 L_{old}^r, S_{hist}^r, 和 X^r，以及数值 L_{old} 和 p_{good}，并使其可供获取。此外，未来工作需要建立自主工作流，以便用户友好地开发 lMLP，并使其能够直接连接电子结构和原子模拟软件。

3. 计算细节

在本研究的 HDNNP 模型中，每个原子神经网络由一个包含 $n_G = 153$个神经元（扩展等变原子中心对称函数）或 156 个神经元（原子中心对称函数）的输入层、三个隐藏层（分别包含 $n_1 = 102$、$n_2 = 61$ 和 $n_3 = 44$ 个神经元）以及一个输出神经元组成。权重初始化方法详见补充信息的 S1.1 节。在预测中，采用了集成规模 $N_{\mathrm{MLP}} = 10$，误差缩放因子为 $c = 2$。需要指出的是，集成规模并未针对具体情况进行优化，但在未来的应用中，可根据数值与不确定性预测的改进效果与额外计算开销之间的权衡进行调整。

对于每个 HDNNP 的训练，我们将参考数据集随机划分为 90% 的训练构象和 10% 的测试构象。在每个训练轮次中，所有训练构象的 10% 用于拟合参考数据集 A 和 B（表2），而 4.07% 用于拟合参考数据集 C。这样，参考数据集 C 每个训练轮次所使用的训练构象数量便与数据集 B 相同。此计数包括自适应选择因子为 $S_{\mathrm{hist}}^r = 0$ 的构象，以及在较晚训练轮次才加入的构象。为表征参考构象的分子结构，eeACSF 向量采用了表1所列的参数。我们使用了大于通常值的 $R_{\mathrm{c}} = 12$ Å 截断半径，以避免静电相互作用方面的问题；静电相互作用并非本文关注的重点。不过，未来的研究可以轻松采用更精细的静电处理方案，例如第四代 HDNNP [28]。

监督训练针对总 DFT 能量减去中性自由原子在最低自旋态下的原子 DFT 能量之和进行（参见补充信息中的表 S1）。所有 S_N2 反应的 DFT 数据均在总电荷为 $-1\,e$、自旋多重度为 1 的条件下获得，其中 e 为元电荷。DFT 计算使用量子化学程序 ORCA（5.0.3 版）[85, 86] 完成。选用 PBE 交换-相关泛函 [87]，搭配 def2-TZVP 基组 [88]，并采用 RI-J 近似。此外，还应用了带 Becke-Johnson 阻尼[90, 91] 的 D3 半经典色散校正 [89]。为平衡参考 DFT 能量和力的训练，损失函数超参数设为 $q = 10.9$。

对于随机梯度下降优化，采用了找到的最佳学习率 0.00075。RPROP 优化采用了——

表1： 径向和角向 eeACSF 分别采用了所列参数的所有组合。所有 eeACSF 的截断半径均设为 $R_c = 12$ Å。除 $H^{\mathrm{ang}} = 1$ 外，均使用了 $\gamma = \pm 1$；对于 $H^{\mathrm{ang}} = 1$，仅采用了 $\gamma = 1$。

径向 eeACSF	
H^{rad}	1, n, m, $\overline{n}$, $\overline{m}$
η^{rad} /Å^{-2}	0, 0.010702, 0.023348, 0.044203, 0.066118, 0.104168, 0.180285, 0.370959, 1.115414
角向 eeACSF	
H^{ang}	1, n, m, $\overline{n}$, $\overline{m}$
η^{ang} /Å^{-2}	0.011238, 0.090144
λ	-1, 1
ξ	1, 2.409421, 9.996864

——参考文献 76 推荐的超参数。例外的是初始学习率，本研究发现 0.001 是最佳选择。对于 Adam优化器，采用了与参考文献 78 相同的超参数，因为未发现通过调整超参数能带来普遍的改进。

CoRe 优化器的超参数为：$\beta_1^{\mathrm{a}} = 0.45$, $\beta_1^{\mathrm{b}} = 0.7$, 0.725, $\beta_1^{\mathrm{c}} = 500$, $\beta_2 = 0.999$,$\epsilon = 10^{-8}$, $\eta_- = 0.5$, $\eta_+ = 1.2$, $s_{\min} = 10^{-6}$, $s_{\max} = 1$,$s_\xi^0 = 10^{-3}$和 $t_{\mathrm{hist}} = 500$。n^m,$_{\mu\nu}$ 冻结 χ 设为权重数量的 1% a^m,$_{\kappa\lambda}{}^{\mu\nu}$ 和 b^m,$_\lambda{}^\nu$ 在各自的族中。例外情况是与输出神经元相关的权重，其 n^m,$_{\mu\nu_{\max}}$冻结 χ 为 0。对于权重 α 和 β, n^m,$_{\mu\nu}$，应用了冻结 $_\chi = 0$。权重衰减超参数 d^m,$_\chi{}^{\mu\nu}$对于 $\chi = \alpha, \beta$设为 0，对于$\nu = \nu_{\max}$设为 0，其余情况设为 0.1。训练进行了 1500、2000 和 2500 个训练轮次。

终身自适应数据集选择的超参数为：$S_{\mathrm{hist}}^{\min} = 0.1$, $S_{\mathrm{hist}}^{\max} = 100$, $\sqrt{T_E^1} = 0.9$,$\sqrt{T_F^2} = 1.2$, $\sqrt{T_F^3} = 2.0$, $N_{F_{--}} = 30$, $N_{F_-} = 100$,$N_{F_+} = 500$, $N_{F_{++}} = 150$, $\sqrt{T_{\mathbf{X}}} = 7.5$, $N_{\mathbf{X}} = 5$,$p_{\mathrm{good}}^{\max} = \frac{2}{3}$, $N_p = 20$和 $\epsilon' = 10^{-6}$。

lMLP 软件使用 Python 编写，并利用科学计算包 NumPy（版本 1.21.6） [92] 以及机器学习框架 PyTorch（版本 1.12.1） [93]。该软件与参考数据集、本研究生成的输出，以及用于分析和绘制这些输出的脚本一同发布在 Zenodo 上（数字对象标识符：10.5281/zenodo.7912832）。

为了公平、可靠地比较不同的多层感知机训练，我们针对每种设置进行了 20 次训练，在划分训练数据和测试数据、权重初始化以及训练数据选择过程中使用不同的随机数。对于每种设置，我们选取表现最好的十个机器学习势进行分析，以减少对具体随机数的依赖。这样可以最大限度地降低异常值的影响，这些异常值可能源于不利的初始参数选择，而这可能影响所有优化器。因此，所提供的均值和标准
偏差表示表现良好的情况。

由于本研究在训练集成成员时采用了不同的训练数据和测试数据划分，因此对参考数据进行的集成预测混合了训练数据和测试数据的预测结果。虽然这种方法高效地利用了所有参考数据进行训练，但仍需借助额外数据对集成进行无偏验证。

4. 结果与讨论

4.1. 参考数据

为比较原子中心对称函数和扩展等变原子中心对称函数的性能，我们构建了参考数据集 A，其中包含十种不同的气相 S_N2 反应。这些反应由 2026 个参考构象表示，其相应的能量和原子力分量由参考方法获得。参考构象对应于不同的化学结构，这些结构涵盖了 S_N2 反应的构象空间，包括中心甲基碳原子和叔叔丁基碳原子上的离去基团 $X^- = Cl^-$、I^- 以及亲核试剂 $Y^- = Cl^-$、HCC^-、I^- （图 1）。需要指出的是，空间位阻会使叔碳原子的 S_N2 反应具有较高的能垒。该组 S_N2反应仅包含四种不同的元素，因此从计算需求的角度而言，仍可采用原子中心对称函数。随着元素数量增加，原子中心对称函数向量呈组合增长，这妨碍了其在包含更多元素的参考数据中的应用。

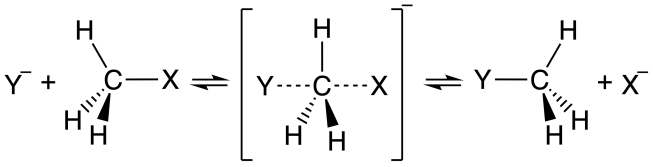


图 1： 甲基碳原子上的 S_N2反应，离去基团为 X，亲核试剂为 Y。

相比之下，eeACSF 向量的大小不随元素数量变化，因此可以训练包含任意数量元素的参考数据。为测试扩展等变原子中心对称函数，我们编制了参考数据集 B，其中包含 8600 个构象，涵盖 42 种不同的 S_N2 反应；离去基团 $X^- =$包括 Cl^-、I^-，亲核试剂 $Y^- =$包括 Br^-、Cl^-、F^-、H_2N^-、H_3CO^-、HCC^-、HO^-、HS^-、HSe^-、I^-、NC^-，中心原子则为甲基和叔丁基碳原子。

参考数据集 A 和 B 均包含通过约束密度泛函理论优化得到的构象。采用不规则采样方法，将碳—离去基团和亲核试剂—碳的距离设定在 1.05 至 5.25 Å 的范围内。距离值的初步网格

针对每个 S_N2反应体系定义了网格值，其中较短距离对应更密集的网格，并为每个体系分别设定最小距离。为获得最终的距离值，对网格值进行了随机调整，并要求这些值不得小于或大于相邻网格值。。

为加入结构畸变的表征，我们通过随机原子位移生成了 12551 个含有中心甲基碳原子的 S_N2 反应体系构象。对于通过约束密度泛函理论优化获得的每个构象，我们在以原子为中心、半径分别为 0.05、0.1 和 0.15Å 的球体内随机移动所有原子的位置，从而生成三个新构象。如果发现某些原子间距离过小，则使用另一组随机位移重新进行该过程。这些构象与参考数据集 B 中的构象共同构成参考数据集 C。需要强调的是，这些参考数据仅用于评估终身学习的性能，可能不足以开展原子模拟，因为可能缺少相关的参考构象。

对参考数据集所做的唯一预处理，是将原子力分量的绝对值限制在最大 15 eV $Å^{-1}$ 以内，以排除仅在极端条件下出现的构象。不同数据集的能量范围、原子力分量范围及标准差汇总于表2，并在以下各节的性能比较中引用。各个 S_N2反应体系的平均能量范围和标准差见补充信息中的表 S4。

表2： 不同参考数据集 A、B 和 C 中的 S_N2 反应数量 N_{S_N2}、构象数量N_{conf}、原子总数 $N_{totalatom}$以及元素 N_{elem}。此外，还给出了能量范围 E^{ref}_{range}、标准差 E^{ref}_{std}、原子力分量范围 $F^{ref}_{\alpha,n,range}$ 以及标准差 $F^{ref}_{\alpha,n,std}$。

参考数据集	A	B	C
N_{S_N2}	10	42	42
N_{conf}	2026	8600	21151
N^{total}_{atom}	22983	100117	190065
N_{elem}	4	10	10
E^{ref}_{range} /毫电子伏/原子$^{-1}$	1715.3	1855.0	2018.0
E^{ref}_{std} /毫电子伏特 原子$^{-1}$	382.6	373.8	391.5
$F^{ref}_{\alpha,n,range}$ / 毫电子伏特 埃$^{-1}$	29883	29984	29998
$F^{参考}_{\alpha,n,}$ 标准 /毫电子伏 Å^{-1}	1098	1066	1706

4.2. 元素包容型原子中心对称函数

对于四种元素，优化后的原子中心对称函数向量（156）与 eeACSF 向量（153）的大小相近（原子中心对称函数的参数见补充信息中的表 S2）。因此，元素数量达到四种时，eeACSF 表示在计算效率上开始占优。表3显示，采用原子中心对称函数和扩展等变原子中心对称函数表示参考数据集 A，得到的 HDNNP 在测试数据上的均方根误差相近，证明了采用扩展等变原子中心对称函数进行结构表示是合理的（集成的预测误差分布见补充信息中的图 S3 (a) 和 (b)）。采用原子中心对称函数的 HDNNP 准确率平均略高。由于不同元素组合的贡献被完全分离，而在扩展等变原子中心对称函数中这些贡献相互混合，因此出现这一趋势并不意外。然而，根据给出的训练和测试均方根误差，扩展等变原子中心对称函数表示不易过拟合。这可能是因为每个扩展等变原子中心对称函数值都取决于截断半径内的所有邻近原子，而原子中心对称函数值仅取决于某些邻近原子。后者或许更适合特定环境，但会降低泛化能力和可迁移性。补充信息中的图 S4 (a)、(b) 和图 S5 表明，采用原子中心对称函数和扩展等变原子中心对称函数表示时，收敛情况和训练过程相似。

表3： 使用原子中心对称函数或扩展等变原子中心对称函数，在参考数据集 A 上训练的单个高维神经网络势（即集成前）和集成模型的均方根误差值。采用 CoRe 优化器和终身自适应数据选择训练 2000年个训练轮次。

单个高维神经网络势	ACSF	元素包容型原子中心对称函数
均方根误差(E^{train}) / meV 原子$^{-1}$	2.0 ± 0.2	2.5 ± 0.3
均方根误差(E^{test}) / meV 原子$^{-1}$	2.3 ± 0.2	2.8 ± 0.3
均方根误差($F^{train}_{\alpha\,n,}$) / meV Å^{-1}	59 ± 3	73 ± 3
均方根误差($F^{test}_{\alpha\,n,}$) / meV Å^{-1}	87 ± 9	92 ± 8
集成		
均方根误差(E) / meV 原子$^{-1}$	0.9	1.3
均方根误差($F_{\alpha,\ n}$) / meV Å^{-1}	34	44

为训练包含十种元素的参考数据集 B，若采用参数η^{rad}、η^{ang}、 λ和 ξ，则原子中心对称函数向量的大小需达到 750，才能获得相同的分辨率。使用向量大小固定为 153 的扩展等变原子中心对称函数，可以避免由此带来的高计算需求。单个高维神经网络势的准确率为均方根误差(E^{train}) $=$ (3.9 $\pm$ 0.4) meV atom^{-1}、均方根误差(E^{test}) = (4.5 $\pm$ 0.6) meV atom^{-1}、均方根误差($F^{train}_{\alpha\,n,}$) = (99 $\pm$ 7) meV Å^{-1}以及均方根误差($F^{test}_{\alpha\,n,}$) = (116 $\pm$ 4) meV Å^{-1}。较高的

表3中结果的准确率较高，原因在于用于训练的数据更少、复杂度也更低，而模型架构和训练超参数保持不变。不过，尤其是高维神经网络势集成的准确率——均方根误差(E) = 2.6meV atom^{-1} 以及均方根误差($F_{\alpha,n}$) = 64 meV Å^{-1}（预测误差分布见补充信息中的图 S6 (a) 和 (b)）——与针对较少元素训练的其他先进机器学习势相当 [19]，这证明扩展等变原子中心对称函数能够表示不同的局部原子环境，包括各种邻近元素。此外，与单个高维神经网络势相比，高维神经网络势集成的准确率显著提高，这说明集成的用途不仅限于不确定性量化。有关高维神经网络势与其他机器学习势的性能比较，请参阅参考文献42和64。

由于待训练的能量范围相对较宽（表2），能量均方根误差略高于通常水平。此外，各个 S_N2 反应体系的平均能量范围也较宽：中心含有甲基碳原子的体系为 747 meV atom^{-1}，中心含有叔-丁基碳原子的体系为 438 meV atom^{-1}。尽管范围较宽，原子力分量的均方根误差仍略低于通常水平，这是因为相当一部分力接近于零。

4.3. 持续韧性（CoRe）优化器

只有当优化器能够在高维参数空间中高效、可靠地找到合适的权重值时，才能得到准确的机器学习势。像随机梯度下降那样使用固定学习率进行训练，会得到准确率较低的 HDNNP。表4中列出的参考数据集 B 的随机梯度下降结果，其大多数均方根误差值几乎比 CoRe 结果大一个数量级，凸显了优化器的重要性。我们发现，对于给定的机器学习模型和参考数据，RPROP 和 Adam优化器是 PyTorch 1.12.1 中性能最佳的优化器。需要指出的是，RPROP 原本适用于一次性对所有数据进行批量学习，而我们使用 RPROP 进行随机优化。RPROP 收敛快速且平滑，但会停滞在不令人满意的准确率上（图 2(a) 和 (b)）。相比之下，Adam优化器需要更多步数才能达到 RPROP 的准确率，但最终能够达到更低的均方根误差值。然而，其收敛过程存在噪声，因而不利于持续应用。

我们的 CoRe 优化器收敛速度甚至更快 RPROP 的收敛效果更好，最终准确率也高于 Adam优化器（图 2 (a) 和 (b)）。其收敛过程仍几乎与 RPROP 一样平稳。因此，CoRe 融合并提升了 RPROP 和 Adam 两者的优点。我们注意到，这些趋势在随机选择训练数据时也成立（见补充信息中的图 S7 (a) 和 (b)），而非采用图 2 (a) 和

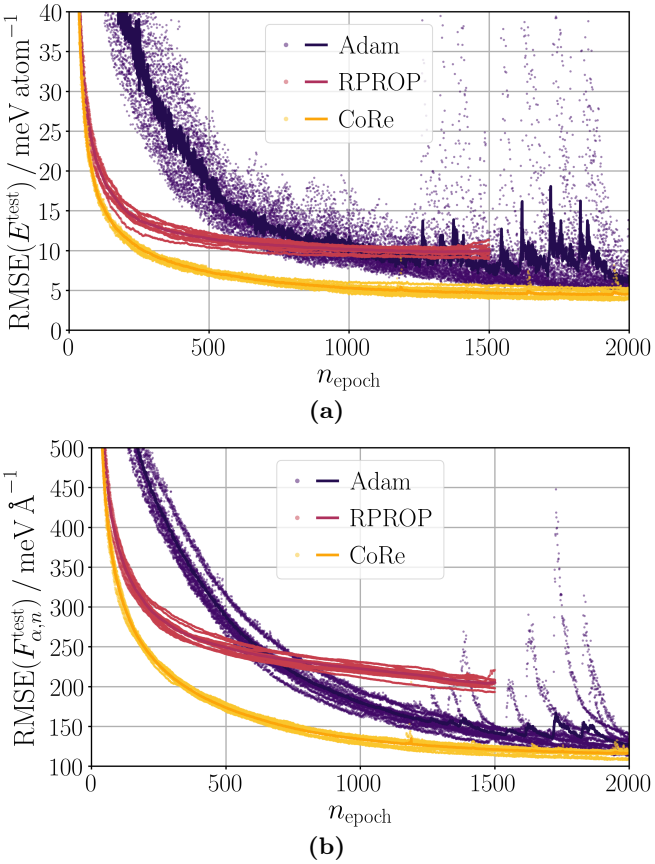


图 2： 使用 RPROP 和 CoRe 优化器训练参考数据集 B 时的收敛情况。图中展示了测试集上 **(a)** 能量 E^{test} 和 **(b)**原子力分量 $F^{\text{test}}_{\alpha,n}$ 随训练轮次变化的均方根误差值 n_{epoch}。单个高维神经网络势的均方根误差值以圆点表示，其均值（不同于集成均方根误差值）以实线表示。优化过程中采用了终身自适应数据选择。

(b).

4.4. 终身自适应数据选择

持续数据集缩减对于终身机器学习至关重要，可在增量学习新数据期间将用于回放的训练数据量控制在可管理的范围内。图3展示了拟合过程中训练数据集如何逐步缩小。训练数据缩减在约 600 个训练轮次后开始，因为要排除数据，必须通过指定次数的连续递减或递增操作，使自适应选择因子 S^r_{hist} 降至 $S^{\min}_{\text{hist}}$ 以下或升至 $S^{\max}_{\text{hist}}$ 以上（算法 2）。使用 RPROP 进行的优化在约 600 个训练轮次后就已趋于平稳。因此，之后的训练轮次中数据排除速度过快，因为大多数数据的损失贡献

表4： 使用优化器 SGD、RPROP、Adam 和 CoRe 在参考数据集 B 上训练的单个高维神经网络势及集成的均方根误差值。优化器和终身自适应数据选择分别应用了 1500（RPROP）和 2000（SGD、Adam、CoRe）个训练轮次。

单个高维神经网络势	SGD	RPROP	Adam	CoRe
均方根误差(E^{train}) /meV 原子$^{-1}$	37 ± 7	9.5 ± 0.7	6.1 ± 1.5	3.9 ± 0.4
均方根误差(E^{test}) /meV 原子$^{-1}$	37 ± 7	10.0 ± 0.7	6.4 ± 1.5	4.5 ± 0.6
均方根误差($F^{\mathrm{train}}_{\alpha\, n,}$) / meV Å^{-1}	564 ± 17	191 ± 5	119 ± 8	99 ± 7
均方根误差($F^{\mathrm{test}}_{\alpha\, n,}$) / 毫电子伏·埃$^{-1}$	557 ± 11	205 ± 8	127 ± 8	116 ± 4
集成				
均方根误差(E) / 毫电子伏/原子$^{-1}$	33	6.8	3.9	2.6
均方根误差($F_{\alpha,\ n}$) / 毫电子伏·埃$^{-1}$	529	131	95	64

构象变化不大，导致自适应选择因子对重要性评估产生偏差。相比之下，采用 Adam优化器进行优化时，收敛行为的波动会导致训练数据缩减缓慢。这是因为数据重要性度量也会发生波动，从而妨碍优化过程越过排除阈值。采用 CoRe 进行优化时，训练数据缩减程度介于 RPROP 和 Adam 之间，因此过程更加均衡。在训练后期的进阶阶段，每个训练轮次排除的构象数量也会减少，使训练更加稳定（另见补充信息中的图 S5）。相比之下，采用 RPROP 进行优化时，由于数据缩减过快且幅度过大，约在 1500 个训练轮次后会变得不稳定。

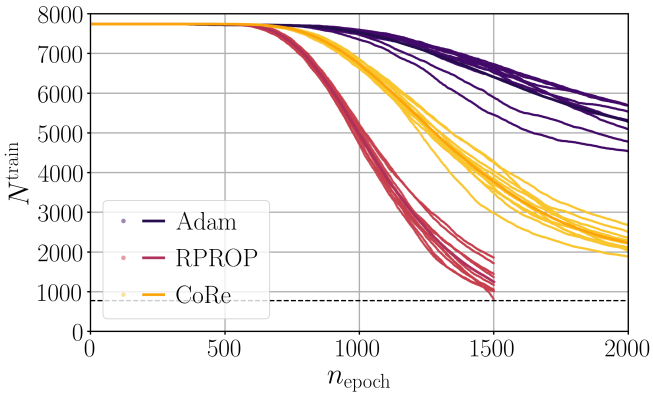


图3： 参考训练数据集 B 中，优化器 Adam、RPROP 和 CoRe 的训练数据集缩减情况。所考虑的训练构象数量N^{train} 随训练轮次n_{epoch}的变化情况如图所示。单个高维神经网络势的 N^{train} 值以圆点表示，其均值以实线表示。黑色虚线表示每个训练轮次中用于拟合的训练构象数量。

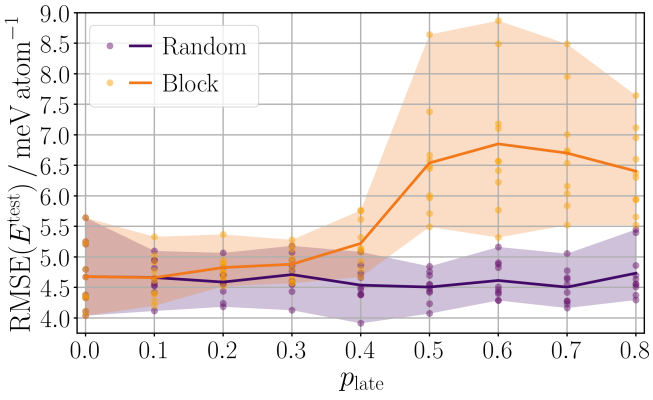
使用 CoRe 训练参考数据集 B 时优化器训练参考数据集 B 时，终身自适应数据选择平均将（5.5 ± 0.3）$\cdot 10^3$ 个训练构象判定为冗余。经过 2000年个训练轮次后，训练数据缩减至初始数据量的 29%。图 2 (a) 和 (b) 表明，这种数据缩减并未导致准确率下降；如果训练使用的是非代表性训练构象子集，准确率本应下降。尽管数据量大幅缩减，与基于全部训练数据进行随机数据选择相比，终身自适应数据选择显著提高了训练准确度（见补充信息中的表 S5 和图 S7 (a) 和 (b)）。所有优化器均呈现这一趋势。对于 CoRe 优化器，与随机数据选择相比，终身自适应数据选择使能量的集成均方根误差值改善了 31%，使原子力分量的集成均方根误差值改善了 48%。

平均而言，有 38 ± 11 个构象因模型无法以较高准确率表示这些构象而被排除在训练之外。这样可以避免这些构象妨碍训练过程。由于在半数以上的训练过程中排除的是相同的 24 个构象，这些构象很可能存在疑问。仅被集成中的少数单个高维神经网络势排除的构象，仍可通过系综平均进行预测（见补充信息中的图 S6 (a) 和 (b)）。这样既能改善单个训练过程，又能保证集成预测的泛化能力。由于某些构象在部分高维神经网络势训练中被排除而产生的不确定性，可通过不确定性量化加以涵盖。因此，这种方法不会影响该方法的可靠性。

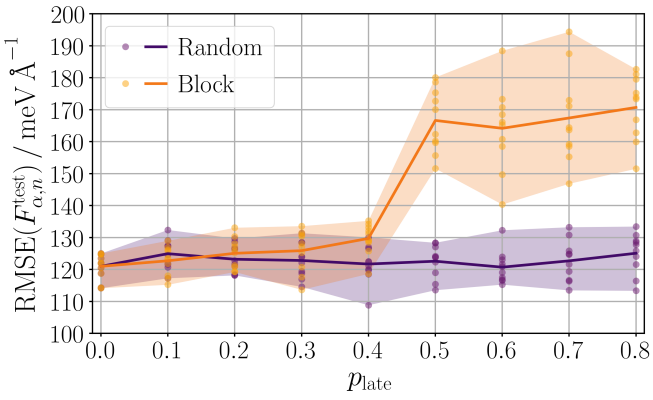
4.5. 终身机器学习势

探讨了三个常见案例，在这些案例中，与数据集扩展的迭代循环相比，终身学习可能更有优势，并且构建——

，并据此训练出使用全部数据的新机器学习势。这些情况分别代表对稀疏采样构象空间的训练数据补全、扩展相同化学体系的表征构象空间，以及学习其他化学体系。终身学习可在每个训练轮次中添加任意数量的新数据点，而且不必像传统的机器学习势主动学习那样采用分块方案。不过，在本研究中，为了清晰地表征由此产生的影响，我们只在单个训练轮次中添加了新数据。



(a)



(b)

图4： 对于 lMLP，在训练后期的某个训练轮次首次提供参考数据集 B 中比例为 p_{late} 的训练数据时的最终准确率。这些数据要么是随机选择的，要么采用特定的数据块。有关该过程的更多详细信息，请参见正文。**(a)** 能量 E^{test} 和**(b)** 原子力分量$F^{\text{test}}_{\alpha,\,n}$ 的测试均方根误差值随晚期数据比例p_{late}的变化情况。单个高维神经网络势的均方根误差值以圆点表示，其均值以实线表示，范围则以较浅色的带状区域表示。采用 CoRe 优化器（$\beta_1^{\text{b}} = 0.7$ “Block”情况下使用 $p_{\text{late}} > 0$，其他情况下使用$\beta_1^{\text{b}} = 0.725$ ）和终身自适应数据选择，训练 1500 轮。

当高比例的随机训练数据在较晚的训练轮次才首次可用时，初始训练轮次中的构象空间采样较为稀疏。

图 4（a）和（b）显示，所提出的终身学习策略能够很好地处理这种情况：最终 lMLP 准确率相对于晚期数据比例 p_{late} 几乎保持不变。对于较高的 p_{late}，加入晚期数据时所处的训练轮次更少，否则少量初始数据会发生过拟合。补充信息中的图 S8（b）和图 S9 显示，额外数据是在测试原子力分量的准确率进入平台期甚至上升时加入的。$p_{\text{late}} \in [0.5, 0.8]$ 的平均集成准确度为 RMSE(E) $= 2.6$ meV atom^{-1}，RMSE($F_{\alpha,\,n}$) $= 68$ meV Å^{-1}，因此与 $p_{\text{late}} = 0$ 的结果非常相似（表4）。除了训练过程更灵活之外，与使用全部数据进行训练相比，终身学习方法在初始训练轮次中所需处理的训练数据更少。补充信息中的图 S9 显示，对于不同的 p_{late}，终身自适应数据选择排除的构象数量在训练 1500 轮后相近。

表 5： 使用参考数据集 C 训练的单个高维神经网络势和集成的均方根误差值；训练采用对包含所有训练数据的静态批次进行学习以及终身学习。在终身学习中，最初 1250 轮仅使用参考数据集 B 的构象进行训练，之后再加入参考数据集 C 的其他构象。采用带有 $\beta_1^{\text{b}} = 0.7$ 的 CoRe 优化器和终身自适应数据选择，训练 2500 轮。

单个高维神经网络势	静态终身 data	学习
均方根误差(E^{train}) / 毫电子伏 原子$^{-1}$	6.1 ± 0.2	6.9 ± 0.6
均方根误差(E^{test}) / 毫电子伏 原子$^{-1}$	6.8 ± 0.2	7.8 ± 0.5
均方根误差($F^{\text{训练}}_{\alpha,n}$) / 毫电子伏·Å^{-1}	168 ± 3	185 ± 12
均方根误差($F^{\text{test}}_{\alpha\,n}$) / meV Å^{-1}	182 ± 4	205 ± 10
集成		
均方根误差(E) / meV atom^{-1}	4.3	4.5
均方根误差($F_{\alpha,n}$) / meV Å^{-1}	121	122

为了考察相同化学体系中表征构象空间扩展情况下的性能，首先使用参考数据集 B 对一个 lMLP 训练 1250 个训练轮次。随后，加入参考数据集 C 中额外的结构扭曲构象，并继续训练 1250 个训练轮次。表 5 显示，终身学习得到的单个高维神经网络势的均方根误差值比在参考数据集 C 的全部训练数据上进行静态批次学习所得结果高约 13%（训练过程见补充信息中的图 S10 (a) 和 (b) 及 S11）。不过，大部分损失的准确

率可由集成模型弥补，因为它能有效降低增加的模型方差（预测误差分布见补充信息中的表 5 和图 S12 (a) 和 (b)）。因此，终身多层感知机概念能够扩展表征构象空间，同时保持准确率。

与参考数据集 B 相比，参考数据集 C 在训练中得到的均方根误差值更高，这是因为其能量范围更大、原子力分量分布也更广（见补充信息中的表2和表 S4），而模型架构和训练超参数保持不变。尽管如此，尤其是集成原子力分量均方根误差，与针对较少元素训练的其他先进高维神经网络势相近[19]。

为研究学习其他化学体系的效率，我们将按体系排序的参考数据集 B 分成两个区块，其中第二个区块所占比例为 p_{late}。构象按照中心碳原子、离去基团和亲核试剂的顺序按字母排序。对于中心叔丁基碳原子，学习其他 S_N2 反应——即在较后的训练轮次加入仅在中心甲基碳原子体系中已知的亲核试剂和离去基团——所得准确率与从一开始就使用全部数据进行学习相近（见图 4 (a) 和 (b) 中的 $p_{\text{late}} < 0.5$ ）。对于 $p_{\text{late}} \geq 0.5$，初始训练数据中仅包含中心甲基碳原子参与的反应，导致最终测试均方根误差值升高。我们强调，对于$p_{\text{late}} \geq 0.8$，初始训练数据中也缺少一些元素。尽管如此，其准确率优于使用 RPROP 得到的结果；在能量方面，则与 Adam 的结果相近（表4）。

与上述情况类似，集成能够有效降低增量学习引入的模型方差，因此是终身机器学习的重要工具。对于 $p_{\text{late}} \in [0.5, 0.8]$，集成准确度为 RMSE($E$) = 3.1 meV atom^{-1}，且 RMSE($F_{\alpha,\ n}$) =78 meV Å^{-1}，因此比 $p_{\text{late}} = 0$ 高约 21%。然而，在这些学习额外化学体系的情况下，终身学习策略仍需进一步改进，才能达到在所有数据上训练的准确度。

由于本研究是针对 lMLP 的概念验证，未来工作需要探索更多不同的化学体系，以微调并改进终身学习策略。尽管如此，我们已表明，终身学习能够达到
与使用全部数据进行训练相比，准确率相同，而训练过程的灵活性则显著提高。

4.6. 集成预测与不确定性量化

为了验证集成预测并考察不确定性量化，将由 CoRe 优化器在参考数据集 B 上训练的 lMLP 应用于通过受限密度泛函理论优化获得的构象，优化过程采用了密集网格——
——该网格覆盖 S_N2 反应中 Br^- +CH_3Cl $BrCH_3$ +Cl^-的 $r_{\text{Cl-C}}$ 和 $r_{\text{Br-C}}$ 距离。我们强调，参考数据集 B 对该反应采用了较稀疏且不规则的采样，因此验证集中的构象不太可能出现在参考数据集 B 中。由此可以验证 lMLP 势能面的平滑性。验证集包含 1062 个构象，其原子力分量最大值为 16 eV Å^{-1}。

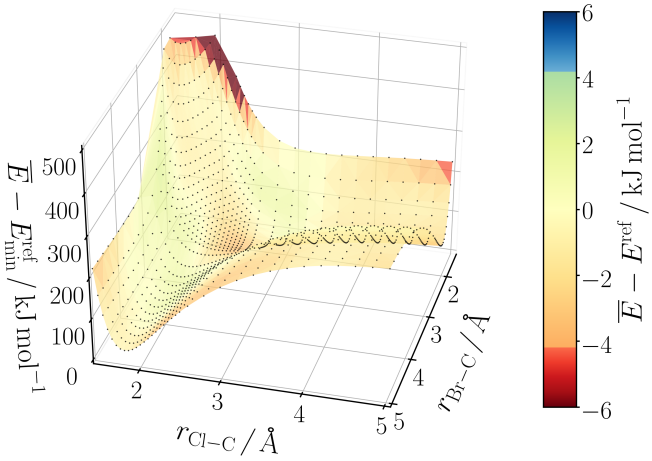


图 5： S_N2 反应 Br^- + CH_3Cl $BrCH_3$ + Cl^-的势能面。lMLP 集成预测能量 E 以给定构象空间中的最低 DFT 参考能量 $E^{\text{ref}}_{\text{min}}$ 为基准；该构象空间由具有受限距离 $r_{\text{Cl-C}}$ 和 $r_{\text{Br-C}}$的 DFT 优化结构所张成。颜色表示 E 相对于 DFT 参考能量E^{ref}的误差。黑点表示 lMLP 的显式评估。黑点之间的颜色为插值结果。

图 5 显示，所示势能面的大多数区域相对于 DFT 参考能量的预测均达到化学准确度，即 1 kcal mol^{-1} = 4.184 kJ mol^{-1}。因此，对于这个六原子体系，误差需小于 7.2 meV atom^{-1}。只有在高能构象下才观察到较大误差，这符合预期，因为最大原子力分量可能比训练数据中的高 1 eV Å^{-1}。训练构象空间内误差较小且分布平滑，证明了 lMLP 的平滑性。

图6 (a) 和 (b) 中讨论了不确定性量化的可靠性。对于不确定性为 $\Delta E \leq 10$meV atom^{-1} 的验证数据，98.6% 的数据点，其不确定性量化值大于或等于相对于 DFT参考能量的绝对误差。对于原子力分量，当 $\Delta F_{\alpha,\ n} \leq 250$ meV Å^{-1} 时，这一比例为 99.7%。因此，对于大多数构象，误差的大小都能得到可靠预测。对于大误差，预计不确定性量化会低估误差（见第 2.6 节），因此上述比例会降至 66%，对于 $\Delta E >$

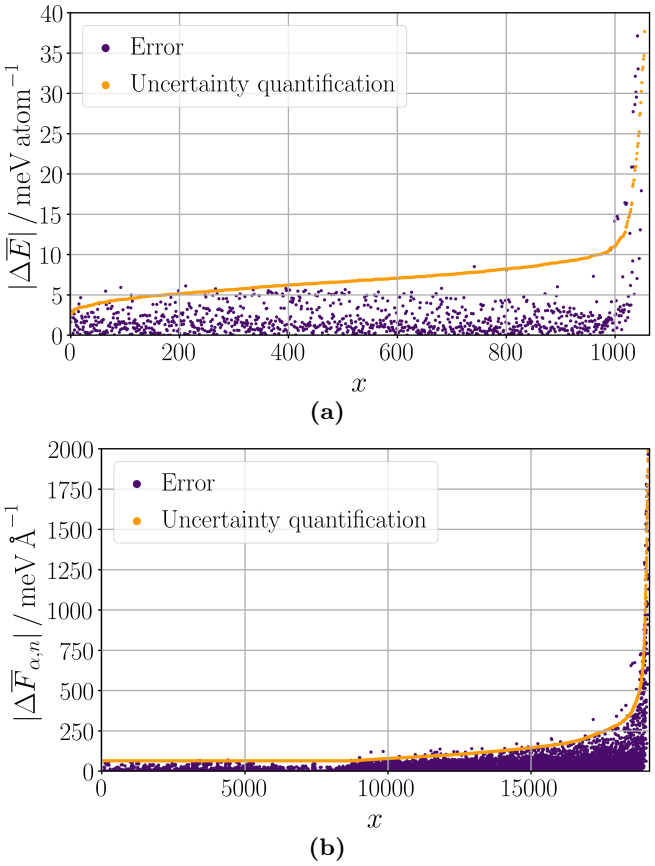


图6： 相对于密度泛函理论参考值的误差绝对值，以及对验证数据集的集成预测进行的不确定性量化，包括 (a) 能量 $|\Delta E|$ 和(b) 原子力分量$|\Delta F_{\alpha,\,n}|$ 。顺序 x 按不确定性量化结果对验证构象进行排序。

10 meV 原子$^{-1}$ 和 93%（$\Delta F_{\alpha,n} \geq 250\,\text{meV Å}^{-1}$）。此外，还能可靠地识别大误差。

5. 结论

本研究提出了 lMLP 的概念，使其能够以滚动方式微调并扩展自身的表征。因此，终身多层感知机概念将多层感知机模型的效率和准确率与灵活性结合起来。对于 lMLP，需要结合通用且计算高效的原子结构表征、多层感知机模型、不确定性量化以及终身学习策略。

因此，我们引入了 eeACSF 向量来表示结构。与许多其他常见的多层感知机描述符相比，它的大小不受化学元素数量的影响。其表示性能在计算成本的盈亏平衡点（约为四种不同元素）与原子中心对称函数向量相近。对于更多元素，由于原子中心对称函数向量的组合增长，eeACSF 向量成为唯一计算上合理的选择。采用扩展等变原子中心对称函数的集成 HDNNP 模型，能够预测包含十种不同元素的 S_N2 参考数据集，其准确率达到以往基于较少元素训练的 HDNNP 所取得的最先进水平。此外，HDNNP 集成是量化模型方差导致的不确定性，以及识别预测不确定性较高的构象的一种可靠方法。此外，集成还能提高准确率，使 S_N2反应的势能面达到化学精度。

作为终身学习策略的基础，我们提出了 CoRe 优化器，它能够结合 RPROP 的快速收敛特性与 Adam优化器的高最终准确率，并加以改进。在本研究的训练过程中，与 Adam优化器相比，CoRe 优化器使 HDNNP 集成在能量和力方面的准确率显著提高了约 33%。采用终身自适应数据选择可进一步提高准确率，并能够缩小训练数据集，在训练过程中剔除存疑数据。CoRe 优化器和终身自适应数据选择也能够改善 lMLP 以外的机器学习模型的训练。

最后，lMLP 可以适应额外的数据，这些数据可以在训练过程的任何阶段持续加入。这样，无需在所有先前数据上重新学习，就能改进 lMLP；同时，由于进行了不确定性量化，所得方法仍然可靠。对于在主动学习过程中获得的数据，或针对相同反应体系扩展构象空间而获得的数据，训练准确率与在包含所有数据的静态批次上学习时相近。借助本文提出的算法，即使加入新的反应体系，准确率也只会适度下降。终身学习的优势在于提升了训练过程的灵活性，使人们能够持续探索化学反应性，并继续训练先前的 lMLP。此外，lMLP 的适应能力对于大型参考数据集尤其有利，因为一次性使用全部数据进行训练在计算上要求很高。我们强调，未来工作中，终身多层感知机概念也可应用于 HDNNP 以外的其他多层感知机方法。

致谢

本研究得到了苏黎世联邦理工学院博士后奖学金的资助。

[1] C. J. Cramer，《计算化学精要：理论与模型》（Wiley，2013）。

[2] F. Jensen，《计算化学导论》（Wiley，2017）。

[3] K. Burke，密度泛函理论展望，《化学物理学杂志》**136**，150901（2012）。[4] C. Riplinger、P. Pinski、U. Becker、E. F. Valeev 和 F. Neese，稀疏映射——用于降低电子结构方法计算复杂度的系统化基础设施。II. 线性标度的基于域的对自然轨道耦合簇理论，《化学物理学杂志》**144**，024109（2016）。[5] S. Das、P. Motamarri、V. Gavini、B. Turcksin、Y. W. Li 和 B. Leback，采用混合精度计算的快速、可扩展且精确的基于有限元的从头算：对金属位错系统进行46千万亿次浮点运算模拟，载于国际高性能计算、网络、存储与分析会议（*SC19*）（美国科罗拉多州丹佛，2019）。[6] A. C. T. van Duin、S. Dasgupta、F. Lorant 和 W. A. Goddard，ReaxFF：一种用于烃类的反应力场，《物理化学杂志A辑》**105**，9396（2001）。[7] S. Piana、K. Lindorff-Larsen 和 D. E. Shaw，蛋白质折叠模拟对力场参数化有多稳健？《生物物理学杂志》**100**，L47（2011）。[8] P. Friederich、F. Häse、J. Proppe 和 A. Aspuru-Guzik，面向下一代物质模拟的机器学习势，《自然·材料》**20**，750（2021）。[9] J. Behler 和 G. Csányi，扩展系统的机器学习势：展望，《欧洲物理学杂志B辑》**94**，142（2021）。[10] J. Behler，展望：用于原子模拟的机器学习势，《化学物理学杂志》**145**，170901（2016）。[11] A. P. Bartók、S. De、C. Poelking、N. Bernstein、J. R. Kermode、G. Csányi 和 M. Ceriotti，机器学习统一了材料与分子的建模，《科学进展》**3**，e1701816（2017）。[12] V. L. Deringer、M. A. Caro 和 G. Csányi，机器学习原子间势：材料科学中崭露头角的工具，《先进材料》**31**，1902765（2019）。[13] F. Noé、A. Tkatchenko、K.-R. Müller 和 C. Clementi，分子模拟中的机器学习，《物理化学年度评论》**71**，361（2020）。[14] J. Westermayr、M. Gastegger、K. T. Schütt 和 R. J. Maurer，关于将机器学习融入计算化学与材料科学的展望，《化学物理学杂志》**154**，230903（2021）。[15] S. Käser、L. I. Vazquez-Salazar、M. Meuwly 和 K. Töpfer，化学中的神经网络势：概念、应用与前景，《数字发现》**2**，28（2023年）。[16] T. B. Blank、S. D. Brown、A. W. Calhoun 和 D. J. Doren，势能面的神经网络模型，《化学物理学杂志》**103**，4129（1995）。[17] J. Behler 和 M. Parrinello，高维势能面的广义神经网络表示，《物理评论快报》**98**，146401（2007）。[18] J. Behler，用于大型分子和凝聚态系统反应模拟的第一性原理神经网络势，《德国应用化学》**56**，12828（2017）。[19] J. Behler，四代高维神经网络势，《化学评论》**121**，10037（2021）。[20] K. T. Schüett、F. Arbabzadah、S. Chmiela、K. R. Müller 和 A. Tkatchenko，深度张量神经网络带来的量子化学洞见，《自然·通讯》**8**，13890（2017）。[21] O. T. Unke 和 M. Meuwly，PhysNet：一种用于预测能量、力、偶极矩及……的神经网络部分电荷，《化学理论与计算杂志》**15**, 3678 (2019). [22] O. T. Unke, S. Chmiela, M. Gastegger, K. T. Schütt, H. E. Sauceda, and K.-R. Müller, SpookyNet：利用电子自由度和非局域效应学习力场，《自然·通讯》**12**, 7273 (2021). [23] I. Batatia, D. P. Kovács, G. N. C. Simm, C. Ortner, and G. Csányi, MACE：用于快速、准确力场的高阶等变消息传递神经网络，arXiv:2206.07697 [stat.ML] (2022). [24] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, 高斯近似势：无需电子即可达到量子力学的准确率，《物理评论快报》**104**, 136403 (2010). [25] A. V. Shapeev, 矩张量势：一类可系统改进的原子间势，《多尺度建模与仿真》**14**, 1153 (2016). [26] N. Artrith, T. Morawietz, and J. Behler, 多组分体系的高维神经网络势：在氧化锌中的应用，《物理评论B辑》**83**, 153101 (2011). [27] S. A. Ghasemi, A. Hofstetter, S. Saha, and S. Goedecker, 基于神经网络获取的电荷密度、具有密度泛函准确率的离子体系原子间势，《物理评论B辑》**92**, 045131 (2015). [28] T. W. Ko, J. A. Finkler, S. Goedecker, and J. Behler, 第四代高维神经网络势：实现准确静电学并纳入非局域电荷转移，《自然·通讯》**12**, 398 (2021). [29] N. Artrith and J. Behler, 金属表面的高维神经网络势：铜的原型研究，《物理评论B辑》**85**, 045439 (2012). [30] J. Behler, 构建高维神经网络势：教程式综述，《国际量子化学杂志》**115**, 1032 (2015). [31] E. V. Podryabinkin and A. V. Shapeev, 线性参数化原子间势的主动学习，《计算材料科学》**140**, 171 (2017). [32] N. Bernstein, G. Csányi, and V. L. Deringer, 势能面的从头探索与自引导学习，《npj计算材料》**5**, 99 (2019). [33] M. Eckhoff and J. Behler, 利用自旋依赖的原子中心对称函数构建磁性体系的高维神经网络势，《npj计算材料》**7**, 170 (2021). [34] T. Morawietz, A. Singraber, C. Dellago, and J. Behler, 范德华相互作用如何决定水的独特性质，《美国国家科学院院刊》**113**, 8368 (2016). [35]M. Hellström and J. Behler, 水溶液氢氧化钠中浓度依赖的质子转移机制：从受体驱动到供体驱动，再反向转变，《物理化学快报》**7**, 3302 (2016). [36] B. Cheng, E. A. Engel, J. Behler, C. Dellago, and M. Ce- riotti, 液态水和固态水的从头算热力学，《美国国家科学院院刊》**116**, 1110 (2019). [37] J. Westermayr, M. Gastegger, M. F. S. J. Menger, S. Mai, L. González, and P. Marquetand, 机器学习助力长时间尺度分子光动力学模拟，《化学科学》**10**, 8100 (2019). [38] S. Amabilino, L. A. Bratholm, S. J. Bennie, A. C. Vaucher, M. Reiher, and D. R. Glowacki, 利用虚拟现实中的交互式量子化学训练神经网络以学习反应势能面，《物理化学杂志A辑》**123**，4486（2019年）。

[39] M. Eckhoff 和 J. Behler，从分子片段到块体：MOF-5 神经网络势的开发，《化学理论与计算杂志》**15**，3793（2019年）。[40] M. Eckhoff、F. Schönewald、M. Risch、C. A. Volkert、P. E. Blöchl 和 J. Behler，利用高维神经网络势弥合锂锰氧化物尖晶石理论与实验之间的差距，《物理评论B辑》**102**，174102（2020年）。[41] M. Eckhoff 和 J. Behler，利用机器学习势深入了解锂锰氧化物—水界面，《化学物理学杂志》**155**，244703（2021年）。[42] Y. Zuo、C. Chen、X. Li、Z. Deng、Y. Chen、J. Behler、G. Csányi、A. V. Shapeev、A. P. Thompson、M. A. Wood 和 S. P. Ong，机器学习原子间势的性能与成本评估，《物理化学杂志A辑》**124**，731（2020年）。[43] J. S. Smith、O. Isayev 和 A. E. Roitberg，ANI-1：一种可扩展的神经网络势，以力场计算成本实现密度泛函理论精度，《化学科学》**8**，3192（2017年）。[44] S. Zhang、M. Z. Makoś、R. B. Jadrich、E. Kraka、K. M. Barros、B. T. Nebgen、S. Tretiak、O. Isayev、N. Lubbers、R. A. Messerly 和 J. S. Smith，利用通用反应性机器学习势探索化学前沿，10.26434/chemrxiv-2022-15ct6-v2（2022年）。[45] D. Hassabis、D. Kumaran、C. Summerfield 和 M. Botvinick，受神经科学启发的人工智能，《神经元》**95**，245（2017年）。[46] Z. Chen 和 B. Liu，*终身机器学习*（Morgan & Claypool 出版社，2018年）。[47] G. I. Parisi、R. Kemker、J. L. Part、C. Kanan 和 S. Wermter，利用神经网络进行持续终身学习：综述，《神经网络》**113**，54（2019年）。[48] D. Maltoni 和 V. Lomonaco，单增量任务场景中的持续学习，《神经网络》**116**，56（2019年）。[49] S. Grossberg，自适应共振理论：大脑如何学会有意识地关注、学习并识别不断变化的世界，《神经网络》**37**，1（2013）。[50] I. J. Goodfellow、M. Mirza、D. Xiao、A. Courville 和 Y. Bengio，基于梯度的神经网络灾难性遗忘的实证研究，arXiv:1312.6211v3 [stat. ML]（2015年）。[51] T. L. Hayes、N. D. Cahill 和 C. Kanan，用于流式学习的内存高效经验回放，载于机器人与自动化国际会议（*ICRA*）（加拿大蒙特利尔，2019年），第9769–9776页。[52] D. Lopez-Paz 和 M. A. Ranzato，持续学习的梯度情景记忆，载于 31st 神经信息处理系统大会（*NIPS*）（美国加利福尼亚州长滩，2017年），第6470–6479页。[53] F. Zenke、B. Poole 和 S. Ganguli，通过突触智能实现持续学习，载于 34th 国际机器学习会议（*ICML*）（澳大利亚悉尼，2017年），第3987–3995页。[54] A. A. Rusu、N. C. Rabinowitz、G. Desjardins、H. Soyer、J. Kirkpatrick、K. Kavukcuoglu、R. Pascanu 和 R. Had- sell，渐进式神经网络，arXiv:

1606 04671 4 . v

[ ] 2022

cs.LG ( )。[55] G. I. Parisi、J. Tani、C. Weber 和 S. Wermter，利用双记忆循环自组织实现时空表征的终身学习，《神经机器人学前沿》**12**，78（2018年）。[56] O. T. Unke、S. Chmiela、H. E. Sauceda、M. Gastegger、I. Poltavsky、K. T. Schütt、A. Tkatchenko 和 K.-R. Müller，《机器学习力场》，《化学评论》**121**，10142（2021）。[57] H. J. Kulik, T. Hammerschmidt, J. Sch midt, S. Botti, M. A. L. Marques, M. Boley, M. Scheffler, M. Todor- ović, P. Rinke, C. Oses, A. Smolyanyuk, S. Curtarolo, A. Tkatchenko, A. P. Bartók, S. Manzhos, M. Ihara, T. Carrington, J. Behler, O. Isayev, M. Veit, A. Grisafi, J. Nigam, M. Ceriotti, K. T. Schütt, J. Westermayr, M. Gastegger, R. J. Maurer, B. Kalita, K. Burke, R. Na- gai, R. Akashi, O. Sugino, J. Hermann, F. Noé, S. Pilati, C. Draxl, M. Kuban, S. Rigamonti, M. Scheidgen, M. Es- ters, D. Hicks, C. Toher, P. V. Balachandran, I. Tam- blyn, S. Whitelam, C. Bellinger, 和 L. M. Ghiringhelli，《电子结构机器学习路线图》，电子结构**4**，023004（2022）。[58] J. Behler，《用于构建高维神经网络势的原子中心对称函数》，《化学物理学杂志》**134**，074106（2011）。[59] A. P. Bartók、R. Kondor 和 G. Csányi，《化学环境的表示》，《物理评论B辑》**87**，184115（2013）。[60] F. Musil、A. Grisafi、A. P. Bartók、C. Ortner、G. Csányi 和 M. Ceriotti，《分子与材料的物理启发式结构表示》，《化学评论》**121**，9759（2021）。[61] S. Gugler 和 M. Reiher，《机器学习分子相似性描述符的量子化学根源》，《化学理论与计算杂志》**18**，6670（2022）。[62] F. Scarselli、M. Gori、A. C. Tsoi、M. Hagenbuchner 和 G. Monfardini，《图神经网络模型》，IEEE《神经网络汇刊》**20**，61（2009）。[63] J. Gilmer、S. S. Schoenholz、P. F. Riley、O. Vinyals 和 G. E. Dahl，《用于量子化学的神经消息传递》，载于 34th 国际机器学习会议（*ICML*）（澳大利亚悉尼，2017），第1263–1272页。[64] C. Chen 和 S. P. Ong，《适用于元素周期表的通用图深度学习原子间势》，《自然计算科学》**2**，718（2022）。[65] N. Artrith、A. Urban 和 G. Ceder，《在含有多种物种的组成中高效且准确地对原子能进行机器学习插值》，《物理评论B辑》**96**，014112（2017）。[66] S. Rostami、M. Amsler 和 S. A. Ghasemi，《用于多组分系统机器学习原子间势的优化对称函数》，《化学物理学杂志》**149**，124106（2018）。[67] M. Gastegger、L. Schwiedrzik、M. Bittermann、F. Berzsenyi 和 P. Marquetand，《wACSF——加权原子中心对称函数：作为机器学习势的描述符》，《化学物理学杂志》**148**，241709（2018）。[68] F. A. Faber、A. S. Christensen、B. Huang 和 O. A. von Lilienfeld，《用于通用量子机器学习的基于炼金术与结构分布的表示》，《化学物理学杂志》**148**，241717（2018）。[69] A. A. Peterson、R. Christensen 和 A. Khorshidi，《应对原子级机器学习中的不确定性》，《物理化学化学物理》**19**，10978（2017）。[70] J. S. Smith、B. Nebgen、N. Lubbers、O. Isayev 和 A. E. Roitberg，《少即是多：利用主动学习采样化学空间》，《化学物理学杂志》**148**，241733（2018）。[71] J. S. Smith、B. T. Nebgen、R. Zubatyuk、N. Lubbers、C. Devereux、K. Barros、S. Tretiak、O. Isayev 和 A. E. Roitberg

，以……逼近耦合簇准确率的

迁移学习实现接近耦合簇精度的通用神经网络势，《自然·通讯》**10**，2903（2019）。[72] C. Devereux、J. S. Smith、K. K. Huddleston、K. Barros、R. Zubatyuk、O. Isayev 和 A. E. Roitberg，将 ANI 深度学习分子势的适用范围扩展至硫和卤素，《化学理论与计算杂志》**16**，4192（2020）。[73] G. Imbalzano、Y. Zhuang、V. Kapil、K. Rossi、E. A. Engel、F. Grasselli 和 M. Ceriotti，分子动力学与采样的不确定性估计，《化学物理学杂志》**154**，074102（2021）。[74] M. Reiher，量子化学研究中的分子特异性不确定性量化，《以色列化学杂志》**62**，e202100101（2022）。[75] M. Eckhoff、K. N. Lausch、P. E. Blöchl 和 J. Behler，利用高维神经网络预测氧化态和自旋态：在锂锰氧化物尖晶石中的应用，《化学物理学杂志》**153**，164107（2020）。[76] M. Riedmiller 和 H. Braun，一种加快反向传播学习的直接自适应方法：RPROP算法，载于国际神经网络会议（*ICNN*）（美国加利福尼亚州旧金山，1993年），第586–591页。[77] M. Riedmiller，多层感知机中的高级监督学习——从反向传播到自适应学习算法，《计算机标准与接口》**16**，265（1994）。[78] D. P. Kingma 和 J. Ba，Adam：一种随机优化方法，载于 3rd 国际学习表征会议（*ICLR*）（美国加利福尼亚州圣迭戈，2015年）。[79] H. Robbins 和 S. Monro，一种随机逼近法，《数学统计年刊》**22**，400（1951年）。[80] C. Igel 和 M. Hüsken，改进rprop学习算法，载于第二届国际*ICSC*神经计算研讨会（2000年），第115–121页。[81] E. Heid、C. J. McGill、F. H. Vermeire 和 W. H. Green，表征化学机器学习中的不确定性，10.26434/chemrxiv-2023-00vcg-v2（2023年）。[82] F. Musil、M. J. Willatt、M. A. Langovoy 和 M. Ceriotti，化学机器学习中快速而准确的不确定性估计，《化学理论与计算杂志》**15**，906（2019）。[83] M. L. Paleico 和 J. Behler，一种用于分析机器学习参考数据和描述符的分箱与哈希方法学习势，《机器学习：科学与技术》**2**，037001（2021）。[84] J. Westermayr 和 P. Marquetand，《分子电子激发态的机器学习》，《化学评论》**121**，9873（2021）。[85] F. Neese，《ORCA 程序系统》，《WIREs 计算与分子科学》**2**，73（2022）。[86] F. Neese，《软件更新：ORCA 程序系统——5.0 版》，《WIREs 计算与分子科学》**12**，e1606（2022）。[87] J. P. Perdew、K. Burke 和 M. Ernzerhof，《简明广义梯度近似》，《物理评论快报》**77**，3865（1996）。[88] F. Weigend 和 R. Ahlrichs，《适用于氢至氡、具有分裂价层、三ζ价层和四ζ价层质量的平衡基组：设计与准确率评估》，《物理化学化学物理》**7**，3297（2005）。[89] S. Grimme、J. Antony、S. Ehrlich 和 H. Krieg，《对 H–Pu 这 94 种元素的密度泛函色散校正（DFT-D）进行一致且准确的从头算参数化》，《化学物理学杂志》**132**，154104（2010）。[90] E. R. Johnson 和 A. D. Becke，《分子间相互作用的后 Hartree-Fock 模型》，《化学物理学杂志》**123**，024101（2005）。[91] S. Grimme、S. Ehrlich 和 L. Goerigk，《阻尼函数对色散校正密度泛函理论的影响》，《计算化学杂志》**32**，1456（2011）。[92] C. R. Harris、K. J. Millman、S. J. van der Walt、R. Gommers、P. Virtanen、D. Cournapeau、E. Wieser、J. Taylor、S. Berg、N. J. Smith、R. Kern、M. Picus、S. Hoyer、M. H. van Kerkwijk、M. Brett、A. Haldane、J. Fernández del Río、M. Wiebe、P. Peterson、P. Gérard-Marchant、K. Sheppard、T. Reddy、W. Weckesser、H. Abbasi、C. Gohlke 和 T. E. Oliphant，《使用 NumPy 进行数组编程》，《自然》**585**，357（2020）。[93] A. Paszke、S. Gross、F. Massa、A. Lerer、J. Bradbury、G. Chanan、T. Killeen、Z. Lin、N. Gimelshein、L. Antiga、A. Desmaison、A. Köpf、E. Yang、Z. DeVito、M. Raison、A. Tejani、S. Chilamkurthy、B. Steiner、L. Fang、J. Bai 和 S. Chintala，《PyTorch：一种命令式风格、高性能深度学习库》，载于 33rd 神经信息处理系统国际会议（*NIPS*）（温哥华，加拿大，2019）第 8026–8037 页。